Solvent Effects in Quantum Chemical-based Methods: I. Defined-sector Explicit Solvent in the Continuum Model Approach for Computational Prediction of pK_a and II. Algorithmic Strategies Towards Inclusion of First Solvation Shell Effects.

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Rebecca Abramson

aus

Australien

Promotionskomitee

Prof. Dr. Kim Baldridge (Vorsitz)
Prof. Dr. Jay Siegel
Prof. Dr. Amedeo Caflisch

Zürich, 2013

Acknowledgements

I first met my research advisor, Kim Baldridge, in January 2005, after completing only one year of my Bachelors' degree. This one-month internship sparked what has now been a nine-year relationship, culminating in this thesis work. I cannot imagine a more dedicated and motivational advisor than Kim, and I am so thankful for the interest, advice and support she has shown towards my professional development, both scientifically and personally. I am also thankful to Jay Siegel for the many interesting discussions and insights.

Having had a connection with the University of Zürich now since 2005, I have crossed paths with so many wonderful people. In particular, I'd like to mention Anne Bowen, Celine Amoreira and Yohann Potier, who immediately welcomed me and patiently taught me GAMESS, Laura Berstis, who has been a constant source of inspiration, Mike Packard, whose patience with all things computer/grid related is highly admirable and greatly appreciated and Fitore Kasumaj; all of whom became close friends. I am also thankful for the friendships with so many other members of the 'old crew'; Derik Frantz, Roman Maag, Silvia Rocha and Anna Butterfield. A particular thank you to Oliver Alleman and Jessica Clavadetscher for sharing this last year with me.

All members of the Baldridge group have been a pleasure to work alongside. In addition to the members already mentioned, I'd like to thank Laura Zoppi, Heidi Weber, Tosaporn Sattasthuchana, Löic Roch, Timm Reuman, Limor Shenar Jackson and Roberto Peverati. I am most thankful to the Grid group (Sergio Maffioletti, Riccardo Murri, Mike Packard and Tyanko Aleksiev) for all the computational support. I am also extremely grateful for the fantastic administrative help offered by our secretarial staff in the OCI (Salomé Fässler, Sarah Amman and Marianne Grima). In particular, I thank Salomé Fässler for following my PhD process from the initial visa application, almost through to graduation, and becoming a very close friend with whom I shared many great moments along the way.

I feel fortunate to have been part of the CMSZH graduate school and I'm very appreciative for the additional opportunities they have offered us.

In addition to my friends from the university, I am most thankful to have made so many wonderful friends in Zürich. I feel particularly indebted to the families that have essentially adopted me over the years for the Shabbats and festivals; I thank the Braden-Golays, the Bessermanns and the Haymanns for feeding me and making me feel a part of their families. Thank you to my own parents for their boundless love, support and encouragement to pursue a career abroad. I cherish the trips to the UK, Israel, Italy, Stockholm, Paris and Barcelona that I've shared with my dad over the past years abroad. My sisters, whom I miss daily, I thank for the love, support and ginger cookies that crossed sees. I also thank my Aunt, Julie, for sending emergency supplies of Australian liquorice and other goodies, and for the wonderful times we spent together in New York and Paris.

Finally I would like to thank Kim Baldridge, Jay Siegel, Joe O'Connor and Amedeo Caflisch, for acting as my Promotionskomittee.

ABSTRACT

Chemical, biochemical and catalytic processes occur in environments where the specifics of structure, molecular reactivity and chemical properties are often very different than in an isolated gas phase environment. It is therefore important to have good predictive models that include the solution environment. Continuum models can in principle handle the majority of the bulk effects that depend on ion screening (dielectric constant). In contrast, the "non-electrostatic" or short-range interactions, such as hydrogen bonding or charge transfer, are not accounted for in the continuum model, and various strategies to include these effects are available. One such strategy is the continuum-cluster methodology, which is an implicit-explicit approach, whereby a small number of explicit solvent molecules are included to capture the short-range interactions and the resultant cluster is treated with a continuum model to capture the long-range or bulk energetics. This thesis work focuses on elucidating a strategy to systematize the number and placement of the explicit solvent molecules included in the cluster for modeling solution phase properties, in particular, dissociation constants. A new model, the Defined-Sector Explicit Solvent in Continuum Cluster Model (DSES-CC), provides a systematic basis for the inclusion of explicit solvation within the continuum model ansatz, resulting in a transferable explicit solvent arrangement for all systems containing a carboxylic or carboxylate moiety. The DSES-CC model achieves benchmark accuracy for prediction of first and second dissociation constants (pK_a^1 and pK_a^2 values) of carboxylic acids. Explicit solvation is shown to be essential for accurate prediction of dissociation constants particularly due to the sensitivity of the property to small changes in free energy of dissociation. All calculations carried out in the development and implementation of DSES-CC have been done with COSab, the locally modified version of the COSMO in GAMESS software package.

While the derived DSES-CC model provides a rigorous means to include first solvation shell effects, optimal use of such ideas would involve an integrated approach, where any functionality could be treated without having to identify a new

DSES-CC for each functional group. In this work, the idea of a distance dependent dielectric function is investigated, where the distance function is dependent on the electron density of the solute system. A number of algorithmic steps towards this goal have been pursued in this work, including a) cavity construction components, b) outlying charge error correction schemes, and c) general efficiency of model algorithmic components. The basic cavity construction routine is enhanced to include a variety of vdW radii options, enabling greater flexibility and user control. The distributed multipole outlying charge scheme is parallelized to achieve an order of magnitude speedup, facilitating the use of the scheme for advanced solvent methodology. An isodensity cavity routine is implemented, which subsequently requires a new tessellation scheme to integrate with distance dependent dielectric models. Consideration of other contributions for achieving chemically significant free energies in solvent phase are also discussed, such as statistical thermodynamics, to encourage stepwise corrections to the electrostatic solvation energy obtained from COSab calculations.

ZUSAMMENFASSUNG

Chemische, biochemische und katalytische Prozesse treten in Bereichen auf, in denen die Eigenschaften der Strukturen, molekularen Reaktivitäten und chemischen Eigenschaften sich oft von denen in einem isolierten Gasphasenbereich unterscheiden. Folglich ist ein gutes vorhersagendes Model wichtig, welches die Umgebung in Lösung mit einbezieht. Im Prinzip können Kontinuum-Modelle die Mehrheit der Ensemble-Effekte (bulk effects) bewältigen, die von Ionen Screening abhängen. Im Gegensatz dazu erklärt das Kontinuum-Modell die nicht elektrostatische Wechselwirkungen oder solche über kurze Distanzen, wie Wasserstoffbrücken oder Ladungsübertragungen, nicht. Es sind jedoch mehrere Strategien vorhanden um diese Effekte miteinzubeziehen. Eine solche Strategie ist die Kontinuum-Cluster Methodologie, eine implizit-explizite Vorgehensweise, wobei eine kleine Menge expliziter Lösungsmittelmoleküle einbezogen werden um Wechselwirkungen über kurze Distanzen zu erfassen und das daraus resultierende Cluster als Kontinuummodell zu behandeln um die weitreichende oder Gesamt- Energetik zu erfassen. Diese Arbeit erläutert eine Strategie der Systematisierung der Anzahl und Platzierung expliziter Lösungsmittelmoleküle, welche im Cluster zur Modellierung der Eigenschaften in der Flüssigphase, insbesondere Dissoziationskonstanten, miteinbezogen werden. Ein neues Modell, Defined-Sector Explicit Solvent in Continuum Cluster Modell (DSES-CC), liefert eine systematische Basis um die explizite Solvatation innerhalb des Kontinuum Cluster Modell Ansatzes miteinzubeziehen, was zu einer übertragbaren expliziten Lösungsmittelanordnung für alle Systeme, welche Carbonsäuren oder Carboxylatkomponenten beinhalten, führt. Das DSES-CC Modell setzt neue Massstäbe für die präzise Voraussagung der ersten und zweiten Dissoziationskonstanten (pK_a^1 and pK_a^2 Werte) von Carbonsäuren. Es wurde gezeigt, dass die explizite Solvatation für die akkurate Voraussagung von Dissoziationskonstanten essentiell ist, insbesondere aufgrund der Sensibilität der freien Energie der Dissoziation für kleine Veränderungen. Alle Berechnungen, die in der Entwicklung und Umsetzung des DSES-CC ausgeführt wurden, wurden mit COSab erstellt, einer lokal modifizierten Version von COSMO im GAMESS Programmpaket.

Das abgeleitete DSES-CC Modell bietet ein ausführliches Hilfsmittel um die Effekte der ersten Solvathülle miteinzubeziehen. Ein optimaler Gebrauch dieser Idee würde einen Ansatz beinhalten, bei dem jegliche Funktionalität behandelt werden könnte, ohne zuerst ein neues DSES-CC für jede funktionelle Gruppe identifizieren zu müssen. In dieser Arbeit wird die Idee einer distanz-abhängigen, dielektrischen Funktion, bei der die Distanzfunktion von der Elektronendichte vom System der gelösten Substanz abhängt, untersucht. Mehrere algorithmische Schritte wurden zur Verfolgung dieses Ziels in dieser Arbeit nachgegangen, unter anderem a) die Komponenten der Kavitätskonstruktion, b) die Fehlerkorrekturen der ausserhalb liegenden Ladung, und c) die Komponenten der generellen Effizienz des algorithmischen Modells. Die grundlegende Routine der Kavitätskonstruktion wird mit der Option eine Vielfalt von vdW radii miteinbeziehen zu können verstärkt, was eine grössere Flexibilität und Benutzerkontrolle ermöglicht. Das Modell der ausserhalb liegenden, verteilten Ladungen wird parallelisiert, um eine Beschleunigung von einer Grössenordnung zu erhalten, was den Gebrauch dieses Modells für eine erweiterte Lösungsmittelmethodik erleichtert. Die Routine der gleich-dichten Kavität wird implementiert, darauffolgend wird ein neues Parkettierungsschema benötigt, um es mit dem distanz-abhängigen Modell zu integrieren. Die Berücksichtigung anderer Beiträge um chemisch signifikante Freie Energien in der Flüssigphase zu erhalten werden auch diskutiert, wie beispielsweise die statistische Thermodynamik, um eine schrittweise Verbesserung der aus der COSab Berechnungen erhaltenen elektrostatischen Solvatationsenergie zu fördern.

TABLE OF CONTENTS

1 Introduction

Proton transfer reactions are one of the most ubiquitous chemical reactions. They govern numerous biochemical processes including protein structure and function, enzymatic reactions, and ligand affinities, and are crucial in chemical synthesis, catalytic reactions and countless other reaction processes. The dissociation constant, K_a, is therefore a very significant calculated property. Whilst classical experimental methods to determine pK_a's[1] are not feasible in a number of conditions (e.g. in the presence of salts or extremes in pH), NMR, specifically Homonuclear-Single Quantum Correlation spectroscopy (HSQC), offers the most promising tool to date for experimental pK_a prediction.[1] Computational strategies are desirable for a number of reasons; they offer the possibility of isolating a molecular system that exists in a complex environment, theoretically shouldn't suffer from issues of reproducibility of results, and can offer fast, low-cost methodology. Therefore, there is a broad interest in achieving a transferable computational strategy for pK_a prediction. This work is divided into two contributions. The first major effort is accurate computational prediction of pK_a using ab initio methodology.

Solution phase acid dissociation constants, as a property of the interaction between the solute and the solvent, are intimately related to solution chemistry, providing a method of obtaining the Gibbs free energies of the reaction, and to gauge the strength of hydrogen bonding[2]. In the reverse, the solvent-solute interactions need to be properly understood in order to derive an accurate theoretical framework for achieving pK_a prediction. This symbiotic relationship makes pK_a an ideal property to test the reliability of the underlying solvent model. Therefore, the second component of this thesis work focuses on accurate solvent model algorithmic strategies, taking into account the information gleaned from the computational pK_a investigations.

The computational methodology employed in this thesis work belongs the framework of ab initio quantum chemistry (QM), including Density Functional Theory (DFT). Solution phase calculations with QM methods are achieved with the use of implicit solvation models. The common feature of implicit solvation models is that the solvent

[1] pK_a is logarithmic constant of the acid dissociation constant, K_a.

is treated as a continuum environment with a dielectric constant specific to the type of solvent environment. Self-consistent dielectric continuum models allow the solute and the solvent to interact self consistently until convergence of the energy and gradient of the solute are achieved. Whilst relying on a number of approximations, these models are largely very successful because they provide an accurate treatment of the long-range electrostatic interactions, which account for most of the energetics. However, for highly accurate property calculation, the remaining short-range contributions can be crucial to achieving chemically significant accuracy. The particular solvation method presented in this work, COSab, was developed in the Baldridge group[3] and has in the past been used successfully on small organic molecules to predict free energy of solvation in agreement with available experimental data. With the efforts described in this thesis, the algorithm has been extended and refined in the General Atomic Molecular Electronic Structure Systems (GAMESS)[4], to consider the first solvation shell effects, and also more efficient use of some of the primary option methodologies within the algorithmic control. This effort has involved the optimization, implementation, and application of several strategies for accurate ab initio modeling of molecules in solution in general, and specifically for the determination of accurate pK_a.

In the following chapter, a brief overview is provided of both the experimentally relevant developments in the understanding of solutions, and the theoretical models that have emerged over the last few decades for accurate determination of solution state properties. Chapter 3 provides the background studies conducted for computational pK_a prediction, and presents the model developed in this thesis work to address this problem, and applies to the model to a small dataset. In Chapter 4 the model outlined in Chapter 3 is extended to a significantly larger set of systems to demonstrate the flexibility of the model for different functionality types. The chapter concludes with a detailed analysis of the successes and limitations of the model. Both chapters 3 and 4 resulted in published papers, which are included in these chapters. Chapter 5 discusses some of the challenges relating to the second derivative contributions to solvation theory. The final chapter, Chapter 6, takes a broader look at solvation models to consider opportunities for further improvement and development.

2 Conceptions of the molecular structure of solutions

2.1 A short overview of solvation chemistry

Playing a central role in a plethora of chemical and biochemical processes, solution and dissolution has been a topic of intrigue since the Greek philosophers.[5] The theorems and thumb rules that are now hard wired into an education in chemistry, such as "like dissolves like," the categorization of solvents as 'polar,' 'non-polar', 'protic,' 'aprotic,' the concept of the solvation shells, are all an outcome of the field of modern solution chemistry which essentially emerged around the end of the 19th century.

Foundations of this field are largely credited to the work of Raoult (1830 – 1901), van't Hoff (1852 – 1911) and Arrhenius (1859 – 1927).[5] In 1887, Arrhenius proposed the theory of ionic dissociation, prior to which it was believed that an electric current was required to separate electrolytes in solution.[6] Menschutkin, in 1890, was instrumental in popularizing the view that the solvent can take an active role in the reaction dynamics rather than being merely an inert, 'space-filler,' demonstrated through reactions of trialkylamines with haloalkanes.[5, 7]

These developments were of such importance to chemistry as a whole, that the first Nobel Prize for chemistry was awarded to J.H. van't Hoff in 1901 for his contribution in the area of solution chemistry, specifically for defining osmotic pressure in solutions, and then Arrhenius followed suit in 1903, awarded the Nobel Prize for his work on ionic dissociation.[7] Van't Hoff was particularly instrumental in marrying physics, mathematics and chemistry, and along with Ostwald and Arhenius is responsible for laying the foundations of the field of physical chemistry.[7-8]

Whilst these prior works were instrumental in sparking the field that is now referred to as solution chemistry, the first golden period is said to have taken place from around the 1930s – 1940s.[6] A number of chemists marked this era, including Debye, Huckel and Bjerrum.[8] The work of Debye and Huckel in 1923 on the theory of ionic activities in solution spurred further developments of ideas relating to static and dynamic ion

mobility's and diffusion coefficients.[6] It was around this time that Born also introduced his model for ionic solvation[6] which will be described in Chapter 2.2.

During this period, Eyring and Daniels and Hughes and Ingold made further developments regarding the association between kinetics and solvents, and in 1935 'transition-state theory' emerged, reached independently by Eyring at Princeton and Evans and Polanyi in Manchester.[7] Transition-state theory represents a significant step in maturation of solvation chemistry as it provides a framework to these initial discoveries relating reaction rates and solvent effects, and one that is still of practical utility today.[7]

The discovery of nuclear magnetic resonance (NMR) and solution X-ray (XD) and neutron diffraction (ND) around the 1950s accelerated the field and transformed the field to one that resembles our understanding of solute-solvent interactions today. Thus the period is referred to as the renaissance of solvation chemistry and it was the first time that the concept of molecular structure was finally applied to understand solutions.[8] These technologies allowed chemists to probe the static properties of ions, such as hydration numbers, but additionally, it allowed for the dynamics properties, such as the rotational motion and vibrational relaxation, to be studied.[6, 9]

Solvated ions are the simplest prototype to understand solute-solvent interactions and therefore, although this is not intended to be a comprehensive review of the field of solution chemistry, this short history frames the context of this work. Ionic hydration is as pervasive as proton transfer in chemical processes, including electrochemistry, surface and membrane phenomena, conformational equilibria in peptides, proteins and nucleic acids and enzymatic processes.[10] Fast-forwarding over a century of developments, the chemical community progressed from believing that an electric current was necessary to overcome the coulombic interaction between ion pairs, to realizing that the interactions with the solvent actually induced this separation, to understanding that solvent also had molecular structure, which led to an explosion of experimental studies to measure hydration numbers and explore the dynamic nature of solvent-solute interactions.

Theoretical chemistry has also played an important role in elucidating how reaction mechanisms in solutions.[11] Essentially two avenues for modeling the role of solvent exist; explicit methods and implicit methods. Fully explicit approaches are limited to the realm of classical mechanics; Molecular Mechanics (MM), Monte Carlo (MC) and Molecular Dynamics (MD) simulations. In contrast, the models available in the Quantum Mechanics (QM) sphere depend generally on an implicit approach that is originally based on Born's continuum solvent model, which was developed in 1920. The focus in this work is QM methods, and the following section provides an overview of implicit solvation models from Born's model to the modern day methodology. Until the advent of computers in the 1950s quantum mechanics remained in infancy. The advances in technology encouraged significant developments in the theory and accuracy of the approximations of the Schrödinger equation. However it wasn't until the 1990s, when Density Functional Theory (DFT) was established, that quantum mechanics could really be used for applications.[12] The use of QM for solvated systems evolved around the same time, and as such, there remain some serious challenges in this field, and therefore also many opportunities for contribution and development.

2.2 Implicit solvation models

In 1920, Max Born first introduced a model to calculate the energy of solvation of monovalent ions. Born's formula calculates the coulombic interaction between a monovalent ion in a spherical cavity and the solvent, treated as continuum with given dielectric constant, represented by,

$$\Delta G_S^{ion} = -\frac{\varepsilon_S - 1}{\varepsilon_S}\frac{Q^{ion^2}}{2R^{ion}} \qquad (2.2\text{-}1)$$

Where R^{ion} is an effective ion-radius, ε_S is the dielectric of the solvent, and Q^{ion} is the total charge on the ion.[12]

This basic electrostatic model, although a crude approximation, was greatly appealing as a complimentary method to the early experimental work on solvation of ions. Very quickly it was developed further to include molecules with higher electrostatic

moments, i.e. dipole, quadrupole, octupole etc., by Kirwood and Onsager.[12] The general formula, as derived by Kirwood, for electrostatic multiple moments is given by,

$$\Delta G_S^X = -\frac{1}{2}\sum_{l=1}^{\infty} f_l(\varepsilon_S)\frac{M^{X^2}}{R^{X^{2l-l}}} \qquad (2.2\text{-}2)$$

with,

$$f_l(\varepsilon_S) = \frac{\varepsilon - 1}{\varepsilon + x_l} \qquad (2.2\text{-}3)$$

and

$$x_l = \frac{l-1}{l} \qquad (2.2\text{-}4)$$

The factor of ½ is a consequence linear response theory, as half the interaction is required for the generation of the response.[12] If truncated at l=1, Kirwood's formula is identical to Born's formula for ions.[12] The expression truncated at l=2, i.e. at dipole moment, gained the most currency, at that time, because experimental methods were not able to provide values for higher electrostatic multipole moments.[12]

The most important step in the development of continuum models was when they were integrated into a quantum mechanic framework, in the 1980s; this being the distinction between 'modern' continuum solvation models and the earlier work.[12] In modern continuum models, also known as 'self-consistent reaction field methods' (SCRF), the solute, which is described quantum mechanically, polarizes the solvent, which is still described as a continuum dielectric medium, which in turn polarizes the solute itself, until self consistency is reached.[12-13]

The development of molecular shaped cavities then followed suit. With SCRF capability, no further improvements in accuracy could be achieved with the spherical or ellipsoidal cavities, the only options at the time. Simply inadequate for many molecular shapes, they either result in the large amounts of the electronic wavefunction lying outside of the cavity, or large areas of vacuum inside of the cavity. However, the extension to molecular shaped cavities is not trivial, as the reaction field response can no longer be solved analytically, as it can be with Onsager's equation. Numerical methods are therefore required.

A number of solutions to the solvent reaction field (or solvent reaction potential) have been proposed, however the main solvent codes use either a generalized Born (GB) approximation or employ the Poisson (or Poisson-Boltzmann (PB)) equation.[14] Cramer and co-workers' SMx solvent models follow the GB formalism.[15] The GB formulism, whilst being a very fast method, does not provide a rigorous way to take the polarization of the solvent into account, and therefore does so with a number of empirical terms.[12] The PB methods employ various numerical solutions of the Poisson equation for either the volume polarization, P(r), at a position r of the dielectric medium,[13a, 16] or the surface polarization charge density, s(r), on its surface. The later will be the focus of this chapter because it lays the foundation of the COSMO methodology.

Starting with the molecular coordinates, molecular shaped cavities are constructed by interlocking atom-centered spheres of a specified atom specific radius. Cavity algorithms now have sophisticated methods of treating crevices and cusps, which cause significant problems to the solution of the boundary conditions as electrostatic field becomes infinite in those regions.[12] With the surface polarization charge density methods, the surface is then segmented into a number of patches with uniform charge density.

The screening charge distribution can be represented by an m-dimensional vector, σ. Different approaches are taken to calculate the screening charges, which are given by,

$$4\pi\varepsilon\sigma(r) = (\varepsilon - 1)n(r)E^{-}(r) \qquad (2.2\text{-}5)$$

where $n(r)$ is the surface normal vector a point r and $E^{-}(r)$ represents the total electric field at the inner side of the surface at this point.[17] In the next section the approach of Klamt & Schuurmann,[17] the Conductor-like Screening Model COSMO, is detailed as the work done in this thesis builds on a locally modified version, COSab.

Although in theory molecular shaped cavities allow for the electronic wavefunction to be contained in the cavity and not penetrate the solvent space, the choice of radii that the cavities themselves are constructed with, are not self-evident. It probes a more fundamental question of the ideal boundary of interaction between a solute and the solvent. This issue remains unresolved in the sense that different QM packages employ

different radii, defining this boundary at a different point. This issue will also be returned to in more depth in Chapter 6.

A related issue, denoted by Klamt & co-workers as, the outlying charge error (OCE) is an artifact of atom centered radii, which can lead to a tail of a wavefunction penetrating the cavity surface and therefore causing contamination of the reaction field.[18] There have been a number of treatments proposed including an integral equation formalism[19] and a biconjugate gradient technique[20]. Chipman describes the outlying charge error as a problem of volume polarization, proposing a volume polarization distribution method.[13a] Cossi et al. propose a modification to PCM incorporating an implicit volume charge approach to correct for the OCE.[21] Further methods of treating the OCE are described in the next section, as they pertain to the solvent code developed by Baldridge & Co-workers, COSab.[3]

An interesting approach that has thus far not achieved much traction is the idea of using an isodensity surface, or a 'zero-flux' surface, as the cavity boundary. By definition, a 'zero-flux' surface, immediately resolves the issue of outlying charge, as it fits the cavity to where the electronic density essentially becomes zero. However, this may not represent the 'ideal boundary' for the interaction between the solvent and the solute. In fact, amongst the few existing isodensity studies, there has been little consensus on which isodensity contour value is the most appropriate representation of the boundary, with values of 0.001, 0.004 and 0.0004 reported with different implementations.[13b, 22] Additionally, many of the current implementations have often been victim to convergences problems, limiting their general utility.[22a]

Gaussian has two implementations with an isodensity cavity; isodensity polarizable continuum model (IPCM)[23] and self-consistent isodensity (SCI-PCM), which was a later implementation that updates the isodensity surface at every geometry change.[13b] The default isodensity value in SCI-PCM is 0.004 a.u., which has been shown to reproduce experimental liquid molecular volumes.[24] A study on a set of 50 weakly interacting pairs showed that distances and density minima for these systems coincided with a 0.002 a.u. isodensity surface,[25] providing another experimental path to determining an appropriate isodensity contour value to use for cavity construction.

Barone et al. compared the SCI-PCM results to those with the other radii available in Gaussian.[22a] They found that whilst the SCI-PCM results were an improvement over standard Pauling and Bondi radii, a single density value is not appropriate across all systems tested.[22a] SCI-PCM was found to be particularly inadequate for anionic systems, with errors of more than 20kcal/mol being reported.[22a]

Chipman and coworkers have an isodensity surface cavity implemented the SS(V)PE code[26] and has contributed a number of studies regarding what is the optimal isodensity contour value[22b, 26]. Originally Chipman discussed the possibility that there is no one ideal isodensity contour, and that the contour would have to be system dependent.[26] However in studies conducted in both 1998 and 2013, comparing contours of 0.0005, 0.001 and 0.002, he concluded that a contour of 0.001 a.u. provides the best agreement with experiment.[22b, 26]

Assuming an ideal boundary between the solute and the solvent could be established, the next problem regards the region known as the cybotatic region, which is the area immediately surrounding the solute, where bulk behavior breaks down.[27] The short-range interactions between the solute and the solvent are the cause of this divergence from bulk behavior. These interactions include a number of effects, namely, cavitation, exchange-repulsion, dispersion, charge transfer, hydrogen bonding and disturbance of the nearby solvent structure.[13a, 27] The 'first-solvation shell effects' is a term that includes a number of these smaller interactions.

There are a number of existing approaches to include these contributions. A common strategy returns to the idea of radii, and involves selecting cavity radii that best reproduce experimental solvation energies, and hence absorb many of issues of the short-range interactions in the parameterized radii.[22b] The SMx suites, for example, use optimized radii.[27] A disadvantage of such methods is that they obscure the physical meaning of the boundary between the solvent and solute.

SMx, PCM and SVPE also augment the bulk model with specific terms for the short-range interactions.[14, 22b, 28] These terms are often a function of the solvent-accessible

surface area (SASA). The SASA represents an effective area that is in contact with the solute. It is defined by rolling a sphere around the solute surface and tracing a line at the centre of the sphere, as it is moves around the surface. For small solute such as water, the radius of this sphere is taken as approximately half the width of the first solvent shell (i.e. ~1-2 Å for water).[27] Taken to be proportional to the number of averaged interactions between the solute and the solvent, the SASA is invoked in implicit correction methods, such as Cramer & Truhlar's SMx models.[27]

In a recent study, Chipman and co-workers formulate implicit terms specifically to treat dispersion and exchange contributions.[22b] The models are tested for a set of solutes in benzene and cyclohexane, where short-range interactions play a significant role, and the dispersion and exchange contributions may not be of comparable magnitude (therefore won't cancel each other out).[22b] Whilst the authors found the results promising, further developments are still expected to follow.[22b]

A class of approaches that specifically treats the first solvation shell effects, known as explicit-implicit approaches, or the continuum-cluster model as termed by Pliego and Riveros.[29] These involve treating a small number of solvent molecules explicitly and then treating the bulk implicitly. The underlying ideas of a split approach can be traced to Frank and co-workers in 1957[7], long before the evolution of modern continuum models, which involve the QM description of the solute. Frank proposed a split model, dividing the solvent interactions into a first solvation shell layer of highly ordered solvent molecules, directly around the solute, which was at the time a monovalent ion, then a layer or disordered solvent molecules, and finally followed the bulk solvent (Figure 3.2.1).[30] Within the context of the solutions at the time, these changes created more problems that they solved,[7] however one can see the development of the understanding of solution dynamics.

Various three-layer approaches have also evolved including, most famously, the effective fragment potential model from Mark Gordon's laboratory[31] and preceding that, the model of Van Duijnen and co-workers which involved representing the explicit solvent molecules as combination of point charges and atomic polarizabilities[32]. The effective fragment potential model describes the explicit solvent

molecules with an effective potential, and the bulk is treated with the Onsager model.[31] The continuum-cluster model is not so much a three layer approach, but a two layer approach as the explicit water molecules exist with the solute in a cluster and are treated quantum mechanically. The continuum-cluster model forms the basis of the model developed in this thesis work and is discussed in the next Chapter. These other older models may be worth revisiting in in the context of the latest developments in the continuum solvation models.

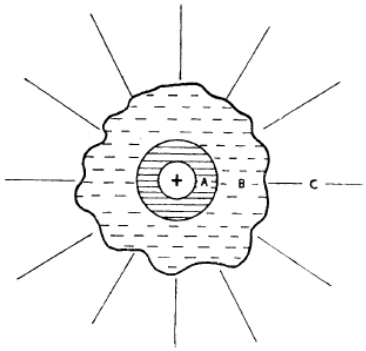


Figure 3.2.1 A model of an ion in solution, taken from Frank & Wen (1957)[30b] with permission from the Royal Society of Chemistry. 'A' represents the region of water molecules with fixed directionality; 'B' represents a region of structure breaking solvent molecules and 'C' is the bulk water.

2.3 Current solvent methods in GAMESS: COSab with features

In this section the specific aspects of the COSMO code, and particularly the developments within GAMESS, are discussed. A major focus of this discussion involves a differentiation of COSMO to the other CSMs, principally in regard to the issues of treatment of the OCE.

COSMO was first developed by Klamt and Schüürman and implemented in to MOPAC.[17] The major difference between Klamt and Schüürman's COSMO and other CSMs is in the solvation of the electrostatic problem; by replacing the dielectric outside the cavity with a conductor, the boundary conditions of the electrostatic

problems are simplified, as the total potential becomes zero at the cavity surface. A scaling factor is then used to correct for the fact that the solvent is not a perfect conductor, i.e. $\epsilon = \infty$, as seen in equation (2.3-1).[33]

$$q = f(\varepsilon)q^* \qquad (2.3\text{-}1)$$

$$f(\epsilon) = \frac{\varepsilon - 1}{\varepsilon + k} \qquad (2.3\text{-}2)$$

Here q is a m-dimensional vector representing the screening charges arising from the polarization of the continuum, which is in turn is due to the m electrostatic potentials, f, defined on the m cavity segments by the charge distribution of the solute, Q.[33] The Coulomb interaction matrix of the screening charges is denoted A.[33] With the condition for a conductor that the total potential arising on the surface segments goes to zero as the solute and the screening charges interact self consistently, then,

$$0 = \phi + A\underline{q}^* \qquad (2.3\text{-}3)$$

The screening charges, q, can therefore be calculated from,

$$q = -f(\epsilon)A^{-1}\phi \qquad (2.3\text{-}4)$$

where equation 2.3-1 has been substituted to find the screening charges, q, in a finite dielectric, rather than the ideal screening charges, q^*.[33]

The total interaction energy of the solute with the screening charges is given by,

$$E_{int} = \underline{\Phi}\underline{q} \qquad (2.3\text{-}5)$$

which is the scalar product of the charges by the total potential arising on the surface segments due to the solute charge distribution, Q.[33] As with Onsager's model, only half of the energy is the response, whereas the other half is due to the creation of the dielectric polarization.[33] The energy resulting from the dielectric screening is given by,

$$E_{diel} = \frac{1}{2}\underline{\Phi}\underline{q} = -\frac{1}{2}f(\varepsilon)\underline{\phi}A^{-1}\underline{\phi} \qquad (2.3\text{-}6)$$

As introduced in section 2.2, the outlying charge error (OCE) is a significant problem pertaining to charge surface models. It is in the calculation of Q that this error arises in the COSMO implementation, because the potential is directly related to the charge distribution of the solute, and therefore if Q is calculated via direct integration, the

wavefunction may not be contained within the cavity. In the original implementations of COSMO in MOPAC and GAMESS,[17, 33] the outlying charge was taken into account by representing the charge density of the solute, Q, as a set of k multipoles, $\underline{M(Q)}$, in addition to calculating Q by direct integration. From the distributed multipole approximation, the potential, $\underline{\phi'}$, which arises on the m segments from the k multipoles, is calculated by,

$$\underline{\phi'} = B\underline{M}(Q) \qquad (2.3\text{-}7)$$

Therefore, B is the (k x m) Coulomb interaction matrix of the multipoles with the segments.[33] It then follows that the interaction energy is calculated by,

$$E_{diel} = \frac{1}{2}\underline{\Phi q} = -\frac{1}{2}f(\varepsilon)B\underline{M}(Q)A^{-1}\underline{M}(Q)B \qquad (2.3\text{-}8)$$

where,

$$-\frac{1}{2}f(\varepsilon)B\underline{M}(Q)A^{-1}\underline{M}(Q)B \equiv \frac{1}{2}QDQ \qquad (2.3\text{-}9)$$

By representing the interaction energy as such, the matrix D is analogous to the coulomb interaction of the charge density Q, which is given by,

$$E^{X}_{Coulomb} = \frac{1}{2}QCQ \qquad (2.3\text{-}10)$$

where C is Coulomb matrix of the solute-solute interactions. This demonstrates that the interaction between a dielectric continuum and a charge distribution is appropriately represented as an additional charge-charge interaction, scaled by the dielectric continuum.[12]

In the reformulation of COSMO by Gregerson & Baldridge in 2003, COSab, introduces a second method to treat outlying charge, along with some other enhancements to make COSMO available to larger molecular systems.[3]

This second method is referred to as the double cavity (DC) method, first conceived by Klamt & Jonas in 1996.[34] This method, as the name suggests, involves building a second cavity a set distance from the first cavity to capture the outlying charge that may lie in this region, and assumes that this accounts for most of the outlying charge. An optimal expansion of 85% (of the original cavity) has been demonstrated by Klamt & Jonas.[34] The segmentation points are mapped one-to-one, and the outer cavity-

segment charges, q', are calculated via $A'q' = -\phi'$, where A' and ϕ' are analogous to A, ϕ, but for the double cavity.[3] Since the segmentation points on the double cavity correspond one-to-one with the primary cavity, the final screening charge is represented as the sum of the two screening charges,[3] i.e.,

$$q'' = q + q' \qquad (2.3\text{-}11)$$

A workflow of the COSab implementation within GAMESS highlighting these two OCE methods is depicted in Figure 2.3.1. Importantly, the double cavity OCE scheme is a post processing procedure, whereas the distributed multiple OCE scheme is computed at every SCF iteration.[3]

Whilst some parameterization has been introduced into COSab via the expansion factor for the double cavity, and in fact through the initial radii that are used for the cavity construction (a topic that is discussed later in chapter 6), the developers have tried to minimize the reliance on parameterized variables. The intention is to facilitate stepwise improvements to the COSab model without double-counting any factors. Currently, COSab does not offer any estimation of the non-electrostatic effects, and therefore provides an ideal basis for the development of schemes that do address these missing contributions.

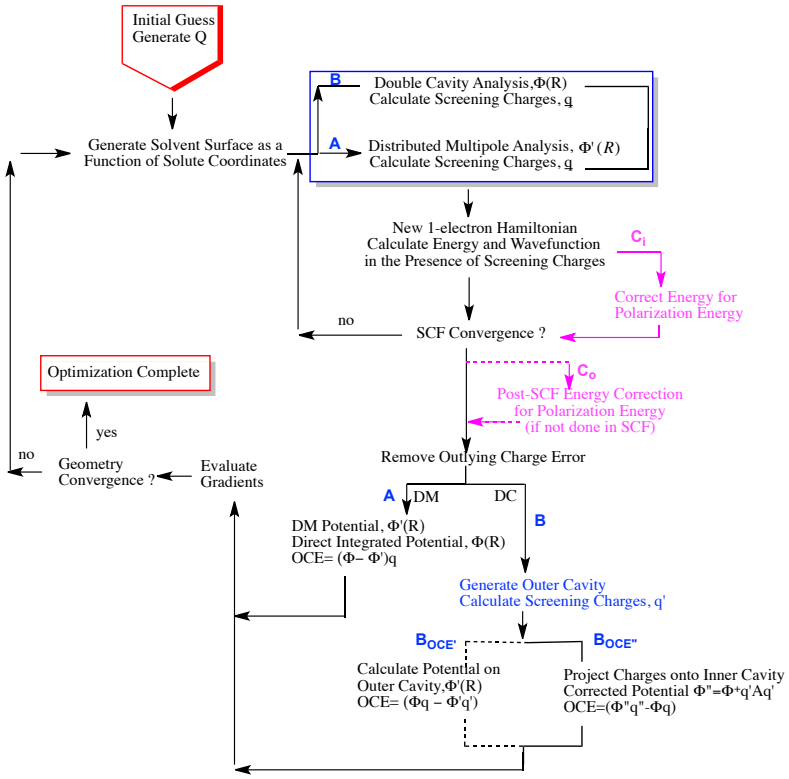


Figure 2.3.1 Workflow of a COSab calculation within the Hartree-Fock-SCF procedure

3 Developing a framework for ab intio pK_a prediction

3.1 Introduction

Acid Dissociation Constants as an aqueous property are challenging to predict computationally because of the inherent sensitivity, of the property to changes in dissociation energy, and consequently, to the inherent deficiencies in continuum solvation methods in obtaining free energies.

This chapter introduces our contribution to the computational prediction of pK_a values, the Defined Sector Explicit Solvent Continuum Cluster (DSES-CC) method.[35] A number of precursory calculations led to the work in this publication, and therefore constitute an important discussion in the progression of this work. These include exploration of a number of existing methodologies in the literature, as well as fundamental QM methodology testing. Figure 3.1.1 provides an overview of all the various theoretical considerations important to calculating pK_a values, providing a neat

framework for the following sections. This includes a discussion of the QM methodology, solvation methodology and the computational strategy of pK_a prediction.

Carboxylic acids were chosen as the test systems in pK_a prediction methodology development for a number of reasons. They fall in the pK_a range of 0 – 5 pK units, which is an ideal range for method development because it suggests strong interaction with the solvent, however not so strong that they cross into the territory of ionic solutions. If one goes to weaker acids, the interactions with the solvent become weaker, which are typically more sensitive calculations. Therefore, carboxylic acids, and specifically, the deprotonated carboxylates, provide an ideal system to study first solvation shell effects.

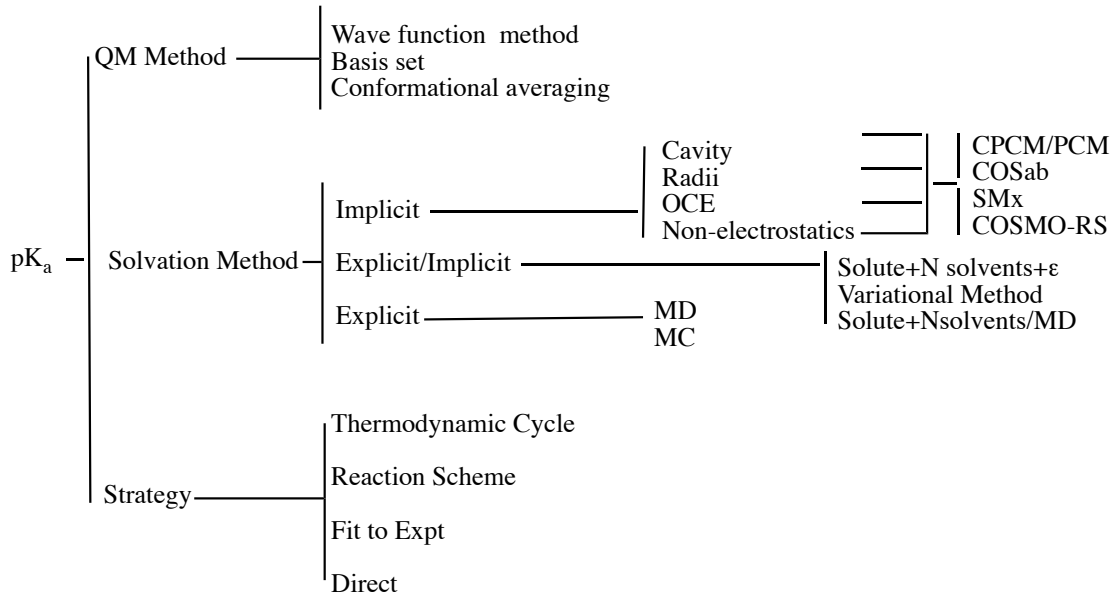


Figure 3.1.1 Tree diagram of the computational methodology considerations for pK_a prediction

3.2 Precursory Calculations

3.2.1 Wavefunction and Basis Set

Wavefunction and basis set are the fundamental building blocks of any QM methodology. We selectively chose B97-D[36], recently implemented and tested in GAMESS[37], as the wavefunction of choice as a compromise in computational time,

and accuracy. Grimme's dispersion corrected density functional, B97-D, has demonstrated close to CCSD(T) level accuracy for non-covalently bound systems[36] and therefore is ideal for this application.

Whilst wavefunction was selectively chosen, a more comprehensive basis set study was carried out (Table 3.2.1-1). Convergence of pK_a value with increasing number of basis functions (valency, diffuse and polarizations) was sought, for two test systems, acetic acid and benzoic acid. These systems were chosen as they represent parent acids for the aliphatic and aromatic classes of carboxylic acids respectively. The full extent of Pople type basis set[38] studies were explored, systematically adding d-type polarization on the heavy atoms, p-type polarization on the light atoms, diffuse functions, increasing the valency, and finally increasing polarization to include f-type on the heavy atoms and d-type polarization on the light atoms. A few Dunning basis sets[39] were also included in the study. Rappoport & Furche's recently developed def2-TZVPPD basis set[40] was included in the basis set study as a final 'gold standard' of modern basis sets. The basis set, a triple-zeta-valence basis set with two sets of polarization and diffuse basis functions, was optimized to calculate polarizabilities.[40] It achieves comparable accuracy to the augmented Dunning or Salej type basis sets, but does not suffer from numerical instability issues typical of augmented basis sets.[40] Furthermore, it shows similar convergence in polarizabilities with density functional and second-order Møller–Plesset wavefunctions.[40]

The basis set study yielded the following observations:

1. Increase from double valent split to triple valent split does not make any notable difference for these systems considered.
2. Increasing polarization beyond (d,p) did not prove beneficial.
3. The most significant difference was seen in the absence of a diffuse function on the heavy atoms. All basis sets without any diffuse functions were anywhere from 4.4 to 14.6 pK units above the experimental value. However, pK_a results, with at least one diffuse function, converged more or less around 6-311+G(d,p), with no improvement being seen with a second diffuse function.

4. Across all standard methods investigated, with at least one diffuse function, the direct values are approximately 1 – 2 pK units from the experimental value.

Although the best results were found with 6-311+G(d,p) for both systems, with prediction 0.81 and 0.73 units of experiment for acetic and benzoic acids respectively we chose the basis set with the extra d polarization, 6-311+G(2d,p). This was chosen to accommodate systems that may require more polarization on the heavy atoms. Furthermore, given that no corrections were included in regard to any of the short-range interactions or other missing contributions one would be wary about picking a result that best matches experiment. Rather the basis set study is important to main requirements in basis set and observe a convergence.

Table 3.2.1-1 Basis set study for acetic and benzoic acids.

	Acetic Acid		Benzoic Acid	
Basis Set	Calc. pK_a	Δ(exptl.-calc.)	Calc. pK_a	Δ(exptl.-calc.)
6-31G(d)	16.97	-12.21	15.26	-11.06
6-31+G(d)	3.12	1.64	2.64	1.56
6-31G(d,p)	19.39	-14.63	17.52	-13.32
6-31+G(d,p)	5.61	-0.85	4.93	-0.73
6-31G(2d,p)	19.91	-15.15	17.32	-13.12
6-31+G(2d,p)	6.18	-1.42	5.53	-1.33
6-311G(d,p)	15.46	-10.70	13.36	-9.16
6-311+G(d,p)	5.57	-0.81	4.98	-0.78
6-311G(2d,p)	15.35	-10.59	12.81	-8.61
6-311+G(2d,p)	6.22	-1.46	5.70	-1.50
6-311G(2d,2p)	15.60	-10.84	13.09	-8.89
6-311++G(2d,2p)	6.53	-1.77	5.58	-1.38
6-311G(2df,2pd)	15.52	-10.76	12.43	-8.23
6-311++G(3df,3pd)	6.86	-2.10	6.22	-2.02
DZV(2d,p)	14.04	-9.28	11.69	-7.49
DZV+(2d,p)	6.46	-1.70	5.87	-1.67
TZV+(d,p)	6.12	-1.36	5.30	-1.10
TZV(2d,p)	9.11	-4.35	8.47	-4.27
def2-TZVPPD	6.38	-1.62	6.12	-1.92

3.2.2 pK_a Strategies

A standard approach to calculate ΔG_{diss} for pK_a prediction invokes the use of a thermodynamic cycle.[41] The thermodynamic cycle involves a two-stage (and consequently a two-method) approach to calculate the change in free energy of a reaction in solution (Figure 3.2.2.1).

Typically, a composite method such as one of the Gaussian-type methods (G*n*)[42] or Complete Basis Set (CBS) methods (e.g. CBS-QB3[43]) is used to compute a ΔG_{gas}, followed by a conventional method for calculation of ΔG_{solv}, as shown in equations (3.2.2-1) and (3.2.2-2).

$$\Delta G_{diss} = G_g(AH) - G_g(A^-) - G_g(H^+) + \Delta G_{solv}(AH) - \Delta G_{solv}(A^-) - \Delta G_{solv}(H^+) \quad (3.2\text{-}1)$$

$$\Delta G_{solv} = G_{aq} - G_g \quad (3.2\text{-}2)$$

The advantage of the thermodynamic cycle is that it results in a cancellation of some amount of error, correcting for some of the systematic limitations associated with gas phase and solution phase computations, and provides a path to obtaining a free energy of dissociation. Different reaction pathways constitute another aspect of strategy, however they are generally combined with discussions of thermodynamic cycle.[44]

The reaction, $AH + H_2O \rightarrow A^- + H_3O^+$, offers another route to accessing the free energy of dissociation, and from a chemical standpoint is considered to be more realistic than having a free proton on the product side. Different thermodynamic cycles have been proposed for this model reaction,[44b, 44c] either using the ΔGs_{olv} or the ΔG_{vap} of H_2O, however given that experimental values are required for both species, H_2O and H_3O^+, this model reaction in practice offers no advantage over, $AH \rightarrow A^- + H^+$.

To briefly illustrate several points, regarding solvent model and thermodynamic cycle, an analysis was carried out again for acetic and benzoic acids. Results were determined using varying cavity radii (UA0, UAHF/UAKS, Pauling, or Bondi), wavefunction type

and basis set, and thermodynamic cycle strategy (none, G3MP2, G3 or CBS-QB3) (Tables 3.2.2-1 – 3.2.2-8). Although only the two acids were studied across these methods, it was considered most unsatisfactory that the only methods to achieve pK_a prediction within 0.74 pK units[2] for both acids were the composite methods coupled with a ΔG_{solv} calculated with Hartree Fock.

Therefore, whilst there is no easy approach to calculating a Gibbs free energy of dissociation without the use of a thermodynamic cycle, the cycle actually obscures clear evaluation of the various error components, and hindering the development of a transferable approach. Consequently, these preliminary studies motivated exploration for a transferable methodology that could be conducted with levels of theory that satisfy the requirements of the systems under investigation.

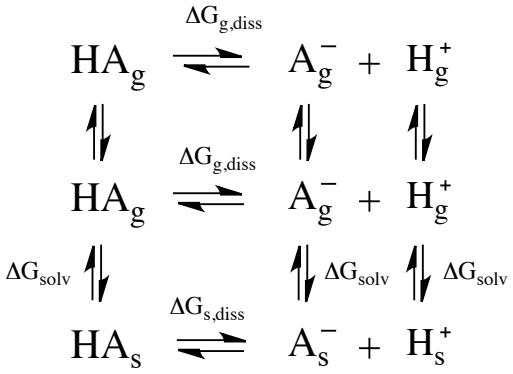


Figure 3.2.2.1 Thermodynamic cycle for a direct deprotonation reaction, involving three levels of calculation; gas phase with a composite method, gas phase with a traditional method, and solvent phase at the same traditional methodology.

[2] The tolerance limit of 0.74 pK units is established in Chapter 3.3. Briefly, it is derived from the fact that 0.74 pK units translates to a difference of 1 kcal/mol in the ΔGdiss.

Table 3.2.2-1 Calculated pK_a values for acetic acid with G3MP2 and specified method for ΔG_{solv}. Experimental $pK_a = 4.76$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	6.16	7.20	11.49	6.67
BMK/6-311+G(d,p)	6.36			
BMK/6-31+G(d)	6.32			
B3LYP/6-311+G(2d,p)	7.10	8.38	12.02	
B3LYP/6-31+G(d)	7.29	8.31	12.03	
HF/6-311+G(2d,p)	4.48	5.56	10.21	
HF/6-31+G(d)	4.53	5.20	10.06	

Table 3.2.2-2 Calculated pK_a values for acetic acid with CBSQB3 and specified method for ΔG_{solv}. Experimental $pK_a = 4.76$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	5.65	6.69	10.98	6.15
BMK/6-311+G(d,p)	5.85			
BMK/6-31+G(d)	5.81			
B3LYP/6-311+G(2d,p)	6.59	7.86	11.51	
B3LYP/6-31+G(d)	6.78	7.80	11.52	
HF/6-311+G(2d,p)	3.96	5.04	9.70	
HF/6-31+G(d)	4.01	4.69	9.55	

Table 3.2.2-3 Calculated pK_a values for acetic acid with G3 and specified method for ΔG_{solv}. Experimental $pK_a = 4.76$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	5.84	6.88	11.17	6.34
BMK/6-311+G(d,p)	6.04			
BMK/6-31+G(d)	6.00			
B3LYP/6-311+G(2d,p)	6.78	8.05	11.70	
B3LYP/6-31+G(d)	6.97	7.99	11.71	
HF/6-311+G(2d,p)	4.15	5.23	9.89	
HF/6-31+G(d)	4.20	4.88	9.74	

Table 3.2.2-4 Calculated pK_a values for acetic acid with no cycle, only specified method for E_{solv}. Experimental $pK_a = 4.76$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	5.10	6.14	10.43	5.60
BMK/6-311+G(d,p)	4.80			
BMK/6-31+G(d)	2.73			
B3LYP/6-311+G(2d,p)	5.72	7.00	10.65	
B3LYP/6-31+G(d)	3.12	4.14	7.86	
HF/6-311+G(2d,p)	8.63	9.71	14.36	
HF/6-31+G(d)	5.18	5.85	10.72	

Table 3.2.2-5 Calculated pK_a values for benzoic acid with G3MP2 and specified method for ΔG_{solv}. Experimental $pK_a = 4.22$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	5.28	6.39	10.54	5.26
BMK/6-31+G(d)	5.16			
B3LYP/6-311+G(2d,p)	6.59	7.50	11.09	
B3LYP/6-31+G(d)	6.60	7.33	11.03	
HF/6-311+G(2d,p)	3.97	4.71	9.17	
HF/6-31+G(d)	3.34	4.25	8.93	

Table 3.2.2-6 Calculated pK_a values for benzoic acid with CBSQB3 and specified method for ΔG_{solv}. Experimental $pK_a = 4.22$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	4.51	5.61	9.77	4.48
BMK/6-31+G(d)	4.38			
B3LYP/6-311+G(2d,p)	5.81	6.72	10.32	
B3LYP/6-31+G(d)	5.82	6.55	10.25	
HF/6-311+G(2d,p)	3.19	3.93	8.39	
HF/6-31+G(d)	2.56	3.47	8.15	

Table 3.2.2-7 Calculated pK_a values for benzoic acid with G3 and specified method for ΔG_{solv}. Experimental $pK_a = 4.22$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	5.16	6.27	10.42	5.13
BMK/6-31+G(d)	5.04			
B3LYP/6-311+G(2d,p)	6.47	7.38	10.97	
B3LYP/6-31+G(d)	6.47	7.21	10.91	
HF/6-311+G(2d,p)	3.85	4.59	9.05	
HF/6-31+G(d)	3.22	4.13	8.81	

Table 3.2.2-8 Calculated pK_a values for benzoic acid with no cyle, only specified method for E_{solv}. Experimental $pK_a = 4.22$

	Pauling	UAHF/UAKS	Default	Bondi
BMK/6-311+G(2d,p)	3.97	5.07	9.23	3.94
BMK/6-31+G(d)	1.59			
B3LYP/6-311+G(2d,p)	5.51	6.42	10.02	
B3LYP/6-31+G(d)	2.95	3.68	7.38	
HF/6-311+G(2d,p)	8.25	8.98	13.45	
HF/6-31+G(d)	4.40	5.31	9.99	

3.2.3 Solvation Method

Solvation method of course is the central theme of this work. Hybrid approaches, referred to as *continuum cluster* (CC), *implicit-explicit* methods, or *super-molecule* approaches, briefly introduced in section 2.2, have gained increasing popularity in addressing pK_a prediction.[45]

The general concept involves addition of a small number (< 3) of solvent molecules positioned close to the solute to account for the first and sometimes second solvation shell interactions. The resulting cluster is then embedded in a cavity and treated with a CSM. While the CC method may have established theoretical grounds, the strategy can be practically challenging due to the need to have a consistent framework for determination of optimal number and position of solvent molecules surrounding the solute. Two general approaches that stand out from the proposed solutions to this problem are highlighted here; regression fit[45a, 45d, 46] and variational method[45c, 47]. In a regression fit strategy, calculated data is statistically fit to experiment (equation 3.2.3-1) in the same manner as has been done for strategies that use a purely implicit solvent system. Typically, when a regression analysis is used, only 1 – 2 solvent molecules are placed around the solute (CC), to provide a consistent approach.[48] Basic thermodynamics ensures that, provided calculated ΔG_{diss} is accurate enough, c_1 should be equal to $\frac{A}{RTln(10)}$ and c_2 should be equal to –log[solvent].[46]

$$pK_a = c_1 \Delta G_{diss} + c_2 \qquad (3.2.3\text{-}1)$$

In these types of regression fits, however, one typically does not find the expected slope of 1/RTln(10), but rather 50-60% of the expected slope.[49] Even when the CC method is used together with a post-DCSM treatment that incorporates a robust statistical mechanics based correction for deviations from basic dielectric behavior and directed interactions (e.g., COSMO-RS) one still finds unexpected nonlinearity.[45a, 46] While the additional post-DCSM treatment provides an improvement over standard DCSMs for pK_a prediction, the nonlinearity of the slope, as well as a noticeable decrease in the overall regression fit as a function of the number of explicit water molecules, are still issues to be addressed.[45a]

The variational method is an alternative method that has been proposed for optimal determination of number and placement of explicit solvent molecules around a solute. Such a method involves minimizing a descriptor, such as ΔG_{solv}, with respect to number and placement of solvent molecules.[29, 45c, 47] For example, Pliego and Riveros[29, 45c] showed the use of a CC scheme for calculation of the solvation free energy and pK_a for a set of ions. Their work signifies an important step towards establishing a criterion for the number of explicit solvent molecules required to represent the first solvation shell and thereby treat different ions in a homogenous manner. Several drawbacks of this method have been pointed out[45a], the most problematic being the unbalanced treatment of species in the reaction considered:
$AH + OH^-.(H_2O)_3 \rightarrow A^-.(H_2O)_n + (4\text{-}n)H_2O$, where only the ionic species is considered within a cluster continuum. More recent versions of the variational method use the total free energy of dissociation as the descriptor, whereby both neutral and charged solutes are solvated with explicit solvent.[47] In either case, one is still faced with the challenges associated with establishing a global minimum cluster representation, and whether or not such a representation is an adequate model for the bulk solvent of a real solvated system.

In the publication to follow, a CC-method is presented that in the broad sense follows a variational principle. The model that is proposed, the Defined-Sector Explicit Solvent in the Continuum (DSES-CC) method, is however the first to provide a rigorous, systematic approach to exploring explicit solvent networks, for the application of pK_a. The basis for network selection is presented in the chapter to follow, however to provide context with the variational principle, the method provides a systematic approach to considering up to five water molecules around the solute, with the objective to finding convergence to experimental pK_a at a fixed degree and configuration of solvation. Furthermore the configuration that reaches this convergence should be the lowest energy configuration at that degree of solvation.

3.3 Defined-Sector Explicit Solvent in Continuum Model Approach for Computational Prediction of pK_a

Authors: Rebecca Abramson and Kim K Baldridge
This work is published in Mol. Phys. (2012) 110(19-20), pp. 2401-2412

3.3.1 Introduction

Accurate prediction of acidity dissociation constant (K_a), is a challenge for the theory of proton transfer reactions. First principles prediction of pK_a within 0.5 pK units of experimental values is a benchmark of broad interest. Considerable efforts have been made in the last decade towards achieving this goal,[46, 50] particularly through the implementation of models using first-shell interactions to obtain accuracy without reliance on fortuitous cancellation of error.[28, 45a, 45c, 45d] In the present contribution, our goal is to investigate in greater detail the relative role of explicit first-shell solvation of carboxylic acids and carboxylates through a defined-sector explicit solvent in continuum cluster (DSES-CC) model that considers the structure-to-chemical affinity relationship of the carboxyl functional group.[51]

Most important to the sector model is a strategy for placement of the explicit solvent molecules with respect to the solute. In order to systematize the study of explicit solvation within the continuum solvation methods, specific solvation states are defined according to degree of solvation and configuration of solvation. The degree of solvation (S_D) is defined simply as the number of explicit solvent molecules implemented. The configuration of solvation (S_C) is defined by the specific set of principal solvation sites (Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}), secondary solvation sites (Q_{br}), and sites within the substituent shell, where solvent has been explicitly implemented. For the particular case of carboxylic acid and carboxylates, a depiction of the principal and secondary explicit solvation sites is illustrated in Figure 3.3.1.1.

The solvation state energy is determined for each component of the acid dissociation reaction ($HA{\bullet}S_C[Q...]$ or $A^-{\bullet}S_c[Q...]$) in a specific S_C within a given S_D. The ΔG of any specific S_C is determined by subtracting the energy of the reactant state from the

product state ($\Delta G = (A^{-}\bullet S_C[Q...] + H^{+}) - HA\bullet S_C[Q...]$). The lowest energy set of S_C within a given S_D (with the lowest energy labelled as S_C^{*}), is used to determine the thermodynamic ΔG_{diss} of acid dissociation for a given S_D ($\Delta G = (A^{-}\bullet S_C^{*} + H^{+}) - HA\bullet S_C^{*}$). The pK_a follows directly as $-\Delta G/2.3RT$,[52] and the calculated value is compared to the experimental value as $\Delta pK = pK_a(exptl) - pK_a(calcd)$. The inherent challenge for QM methods is that at ambient temperature 0.7 kcal/mol error in the ΔG_{diss} misestimates the pK_a by the benchmark 0.5 pK_a unit, whereas ±1.0 kcal/mol accuracies in energy is still a difficult level to achieve using QM solvent strategies.

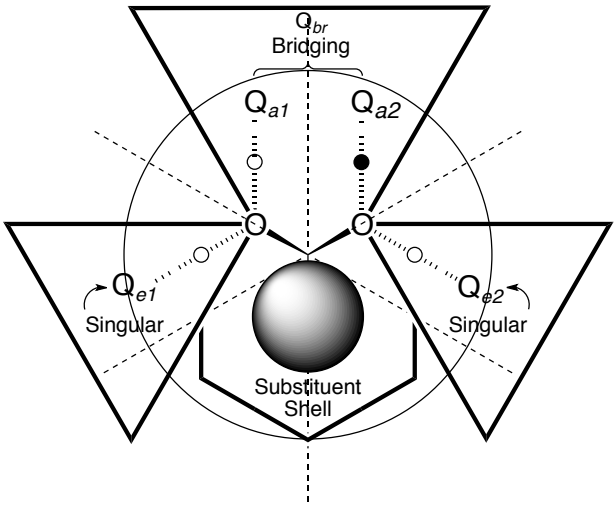


Figure 3.3.1.1 Depiction of the principal and secondary explicit solvation sites around a carboxylic acid (or carboxylate). Small circles indicate presence (filled) or absence (open) of H.

Various contributions to the non-electrostatic term, most importantly cavitation and dispersion-repulsion confound current models. Our strategy focuses on the implicit inclusion of directed effects through the explicit consideration of primary waters of hydration; a key desire being to avoid any added influence of parameterized non-electrostatic terms at this stage. Under the assumption that the differential cavitation term between carboxylic acids and carboxylates is negligible, one should achieve a high level of accuracy if explicit solvation is properly handled. Unique to the defined-sector model approach compared to other explicit solvation studies is a greater depth of analysis involving networks of explicit solvent molecules on the

individual components of the proton transfer reaction. The goal is two-fold in that this model allows a performance assessment of different S_D and S_C levels, and the performance assessment provides a rational basis for the development of a set of solvation networks applicable to new systems.

3.3.2 Computational Methods

All reported calculations were carried out using the GAMESS electronic structure program. The dispersion enabled density functional, B97-D[36] as recently implemented and tested,[37] was determined optimal for the present work. The B97-D functional is a special reparameterization of the original B97 hybrid functional of Becke,[53] An ultrafine grid specification, NRAD=96 NLEB=1202, was used, making the method less susceptible to spurious dispersion contamination in the exchange component. We have previously carried out parameter optimization for several combinations of functional and basis set, including those represented in the present work.[37]

Careful investigation has been made within each of the areas delineated in Figure 3.1.1, including basis set, wave function, thermodynamic cycle, reaction scheme, and solvent radii comparisons, for final determination results (e.g., see SI material). A basis set investigation was conducted using both Møller-Plesset perturbation theory (MP2)[54] and B97-D. Acetic acid and benzoic acid were chosen for the more extensive basis set study, as parent acids for the aliphatic and aromatic classes of carboxylic acids, respectively. The study supported reliability of a triple-z level basis set, 6-311+G(2d,p),[55] for the calculation of the properties in this work. Further polarization and or diffuse functionality (up to 6-311++G(3df,3pd)), resulted in no significant improvements. However, across all standard methods investigated, the direct values without consideration of first solvation shell effects are still approximately 1.5 pKa units from the experimental value. No advantage was found using MP2 compared to B97-D.

Solute-solvent clusters were fully optimized at the B97-D/6-311+G(2d,p) level of theory. Solvation was taken into account using the most recent implementation of our COSab solvation method, based on the original COSMO theory of Klamt and modified for *ab initio* theory within the GAMESS software.[3, 33, 56] Dielectric permittivity of water (ε = 78.4) and of ether (ε = 4.335) were considered in these studies. The parameters of the cavity construction are: 1082 points for the basic grid, 92 segments on the complete sphere. The outlying charge error was taken into account via the double cavity approach.[3, 34] The solvent radial extent was optimized in the parameter optimization studies, and taken as 1.3 for the application studies. Solvent atomic radii were taken from Bondi[57] or from Klamt.[58] All optimizations were performed in the solvent continuum model framework. Analytic Hessian calculations were performed to characterize the structures and determine zero point energy corrections. Final supercluster representations were depicted using MacMolPlt.[59]

3.3.3 Theoretical Approaches for determination of pK_a

Several recent and extensive reviews have been extremely valuable in terms of gathering perspective and general comparison of theoretical approaches for computational determination of pK_a.[45a, 50a, 60] It is not our goal to repeat what has already been well reviewed or discussed in individual articles, but only highlight components relevant to the discussion at hand. In particular, we focus on recent efforts to make predictions within 0.5 pK_a units of experiment,[61] which is the minimum level of accuracy for many problems, such as structure based drug design or synthetic strategies for specially designed ligand complexes. Most strategies benefit from error cancellation schemes and thus arguably suffer from limited transferability due to variation in the error components for variant systems. As a consequence, it can be difficult to maintain accuracy across a large enough data set of molecule to show strong predictability.

Figure 3.1.1 summarizes several of the primary considerations in the design of a computational strategy to achieve this goal. Typical strategies to compensate for

inherent failures in one or more of these areas include statistical fitting to experiment[48a-d, 49, 62] and use of one of many variations in thermodynamic cycle,[44b, 45c, 63] including different reaction schemes.[44b] General efforts to improve the accuracy of the solvent model itself, through hybrid approaches or empirical corrections, have gained increasing attention,[29, 44a, 45c, 47, 48d, 48e, 64] however, they are still typically coupled to strategies to compensate for missing effects of solvation. Additionally, issues such as the importance of statistical thermochemical effects,[61a] and the use of energy-optimized vs. kinetic energy partitioned positioning of solvent molecules, remain debated.

Hybrid approaches such as offered by molecular dynamics simulations in combination with electronic structure methods offer an alternative for determining solvent molecule positioning around solutes;[65] however, in the present work, emphasis is on the performance of fully QM approaches referred to as *continuum cluster* (CC), *implicit-explicit*, or *super-molecule* methods,[45a, 45c, 45d, 47, 50a, 64d] which describe one and the same concepts for general determination of solvation energies and properties, by attempting to capture missing first and sometimes second shell directed effects. The general concept involves addition of a small (< 3) number of solvent molecules positioned close to the solute and the resulting cluster embedded within a continuum. While this strategy may have established theoretical grounds, the implementation can be practically challenging due to the need to have a consistent framework for determination of optimal number and position of the explicit solvent molecules. Very few studies elaborate on the nature of the solvent cluster, and those that do typically concentrate on only a very small set of molecules.[45d, 48e] It is to this end that the present investigation was undertaken, to specifically identify a generalizable working model for explicit solvation, as exemplified for the case of carboxylic acid structure.

In this work, we implement what we term the defined-sector explicit solvent in continuum cluster (DSES-CC) model approach, relying only on solution phase computation (i.e., eliminating use of a thermodynamic cycle) and the use of the sector model for placement of explicit solvent molecules. This use of only solution phase computation is particularly appealing as it eliminates a number of sources of possible error. The approach makes use of the reaction scheme,

$$AH\bullet (H_2O)_n \rightarrow A^- \bullet (H_2O)_n + H^+$$

which avoids complications regarding balancing of the reaction when only the anionic species is considered within a cluster of solvents. Experimental as well as theoretical values have been used for the free energy of the proton in the literature, due to the associated difficulties for determining this quantity directly.[66] In the present work, we use the generally agreed upon value of -265.9 kcal/mol.[44a, 45d, 50a, 67] The gas phase energy is indisputably derived from an enthalpy contribution, 2.5RT, and an entropic contribution calculated from the Sackur-Tetrode equation, yielding a value of -6.28 kcal/mol.[63c]

3.3.4 Results and Discussion

Data arrays from the systematic DSES-CC method were calculated at B97D/6-311+G(2d,p) for the set of carboxylic acids. Although ultimately we are seeking predictability within 0.5 pK units for a specific $S_D(X)$ with HA and A^- in their thermodynamically favored S_C, it is possible that one would find a range of 'acceptable' results due to statistical-thermodynamic factors. However, among any range of potentially acceptable S_C pairs, only those S_C within kT of the thermodynamically favored pair need be considered, as others would not be energetically feasible. A half a pK unit is, in energy terms, only 0.68 kcal/mol, so we define an acceptable range to be within 1 kcal/mol accuracy of the experimental value, or 0.74 pK units. Evaluation across an array of solvation states reveals preferential patterns of limited direct solvation and indicates how various S_C serve for the prediction of pK_a in a specific carboxylic acid.

An important fundamental finding is that determination of pK_a at the level of S_D for carboxylic acid/carboxylate pairs reveals that specific S_C for HA are generally not optimal for A^-. Preferred networks for the neutral versus anionic species tend to be quite different, with the positioning of explicit solvent of the acid being typically more sensitive. This result is in contrast to previous investigations, where consistent positioning of solvent molecules for both the neutral and ionic species is often used.[48e] Identical HA:A^- networks may be found to be transferable for a limited set of

carboxylic acids; however for others, varying functionality and/or additional steric interactions will influence the positions of the explicit solvent molecules differently in the two species. Ultimately, we are looking for convergence across S_D, and general transferability of that S_D with associated S_C(acid):S_C(anion) pair, to an arbitrary system.

To best illustrate the principals of our sector model for explicit solvation, as well as overarching utility, we detail results for several classes of key carboxylic acids. We begin with a more detailed analysis of two specific carboxylic acids, acetic and formic, that serve as the training set, where we introduce degrees of solvation encompassing $S_D(0) - S_D(III)$, including $S_D(III)$ with solvent configurations having a bridging solvent, Q_{br}. With this training set in hand, we move forward to a predictive (test) set of carboxylic acids, where we incorporate 'preferred' solvent configuration(s) identified in the training set to predict S_D for three subsets of carboxylic acids that illustrate a) increase in steric bulk, b) effect of electronic substituent, and c) effect of aromatic substituent, on the basic carboxyl moiety.

3.3.4.1 Training Set: (1) Acetic Acid.

Figure 3.3.4.1.1 and Table 3.3.4.1-1 summarize various possible sector model solvation networks, S_D, for acetic providing an indication of whether the particular configuration of solvation, S_C, is overestimating or underestimating pK_a and to what extent. In addition, the lowest energy S_C of each set is identified as S_C^*, and the HA:A solvent configuration pair where both configurations are lowest energy configurations is highlighted in bold for reference.

In accord with Figure 3.3.1.1, the important solvent positions are Q_{a1}, Q_{a2}, Q_{br}, $Q_{e1,}$ and Q_{e2}. Without any additional explicit solvents, the pK_a for acetic acid is overestimated by 1.44 pK units, well outside our target pK_a range of 0.74 pK units. A single explicit solvent placed at Q_{a2} does not change this significantly. Moving to $S_D(II)$, we illustrate 4 variations in explicit solvent position (Figures 3.3.4.1.1a – 3.3.4.1.1d)): (a) $S_C^*[Q_{a1}Q_{a2}]$:$S_C[Q_{a1}Q_{a2}]$, (b) $S_C^*[Q_{a1}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e2}]$, (c) $S_C[Q_{a2}Q_{e1}]$:$S_C[Q_{a1}Q_{a2}]$, and

(d) $S_C[Q_{a2}Q_{e1}]$:$S_C^*[Q_{a1}Q_{e2}]$, for HA:A pairs, respectively. With one exception (d), all greatly improve the pK_a prediction to within our target pK_a range, ranging from an overestimation of pK_a by 0.18 and 0.54, to an underestimation of pK_a by 0.66. Comparison of (c) and (d) illustrates the sensitivity of the acid towards explicit water. The closest predicted value with respect to the experimental value is obtained with (b) $S_C^*[Q_{a1}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e2}]$, where S_C's in this S_D(II) are the thermodynamically favored HA and A^- structures. Notably, this S_D has differing solvent position networks for the neutral versus anionic species, with the former preference for Q_{a1} and Q_{a2} and the latter preference for Q_{a1} and Q_{e2}, positioning that will reappear in other carboxylic acids.

Moving on to S_D(III), 4 variations of S_C HA:A combinations are investigated: (a) $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, (b) $S_C[Q_{a1}Q_{e1}Q_{e2}]$: $S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, (c) $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C[Q_{a1}Q_{br}Q_{a2}]$, and (d) $S_C[Q_{a1}Q_{a2}Q_{e1}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$. Using an identical networking arrangement for the acidic and neutral form of acetic acid as given by S_C pairs in (d) (Figure 3.3.4.1.1(f)) results in a predicted pK_a that is more than three pK units lower than the experimental value of 4.76. We note a previous study using this same configuration of solvation at the B3LYP/6-311++G(d,p) level of theory provided a pK_a result of 4.86.[48e] However, the distinct differences in DFT type, and in particular, the use of a double thermodynamic cycle in that study, provides sources of inherent error cancellation, leading to result close to experiment. Importantly, the particular solvent network cluster itself is not arbitrary, as shown with this comparison. While it is often more convenient to implement the same solvent cluster for both acidic and anionic species, one should not expect an identical network to solvate each optimally, as the electronic structure of the acid and the anion are quite different. If one modifies the positioning of the identical networks to instead be as in (g), we see a significant improvement in predicted pK_a, at an overestimation of 0.65 pK units. In this case, the offset of the explicit waters through a bridge water molecule establishes a more optimal solvent pacification for the solute. The 'preferred' solvent network is provided by the thermodynamically favored S_C HA:A pair of the S_D(III) set (a), with an estimated pK_a that is underestimated by only 0.18 pK units. Again we see different solvent configurations, S_C, for neutral and anionic species, with the former having a preference for the bridged network, and the latter

having a network that involves Q_{a1}, Q_{e1}, and Q_{e2} to better accommodate the extra negative charge.

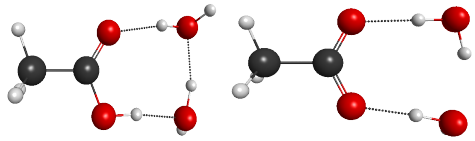
While one may expect to see a convergence to the experimental pK_a at S_D(IV), the results actually are worse (Figure 3.3.4.1.1(i) and (j)). Here, the excess explicit water network unbalances the electronic structure of the solute such that the pK_a is underestimated by well over one pK unit, offering no benefit over a pK_a prediction of just the raw solute. Upon further examination of other possible solvent networks, any combination of water placements that disrupt the "natural" electronic structure of the solute, such as in Figure 3.3.4.1.1(f), (i), or (j), leads to a pK_a result that is much worse than not including any additional explicit solvent molecules at all.

Across all possible degrees of solvation for acetic acid, the observation thus far would indicate that a solvation degree for acetic acid of S_D(II) with $S_c[Q_{a1}Q_{a2}]$:$S_c[Q_{a1}Q_{e2}]$ or S_D(III) with $S_c[Q_{a1}Q_{br}Q_{a2}]$: $S_c[Q_{a1}Q_{e1}Q_{e2}]$ would be reasonable to considered for a 'preferred' transferable network for general use. S_D(II) looks attractive since it only takes two explicit water molecules, but S_D(III) is equally good for prediction of pK_a within our tolerance and both are thermodynamically favored pair sets.

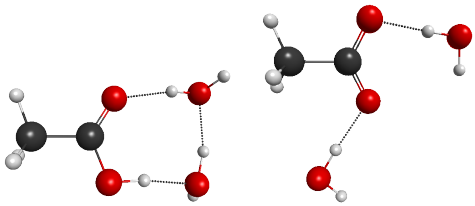
Table 3.3.4.1-1 B97D/6-311+G(2d,p) direct-sector explicit solvent in continuum model results for acetic acid (exptl $pK_a = 4.76$)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
S_D(0)				
$S_c(0)$		8.45	6.20	-1.44
S_D(I)				
HA	**A^-**			
$S_C[Q_{a2}]$	$S_C[Q_{a2}]$	8.49	6.23	-1.47
S_D(II)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{a2}]$	$S_C[Q_{a1}Q_{a2}]$	7.23	5.30	-0.54
$S_C^*[Q_{a1}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e2}]$**	**6.74**	**4.94**	**-0.18**
$S_C[Q_{a2}Q_{e1}]$	$S_C[Q_{a1}Q_{a2}]$	5.58	4.10	+0.66
$S_C[Q_{a2}Q_{e1}]$	$S_C^*[Q_{a1}Q_{e2}]$	5.10	3.74	+1.02
S_D(III)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**6.24**	**4.58**	**+0.18**
$S_C[Q_{a1}Q_{e1}Q_{e2}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$	2.22	1.63	+3.13
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}]$	7.38	5.41	-0.65
$S_C[Q_{a1}Q_{a2}Q_{e1}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$	5.99	4.39	+0.37
S_D(IV)				
HA	**A^-**			
$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$	$S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$	3.19	2.34	+2.42
$S_C^*[Q_{a1}Q_{br}Q_{a2}Q_{e2}]$	**$S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$**	**4.84**	**3.55**	**+1.21**

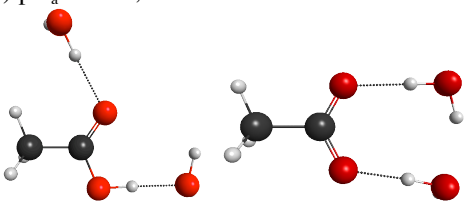
(a) $pK_a = 5.28, \Delta = -0.54$



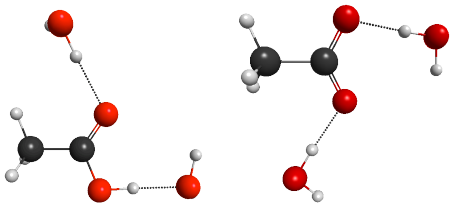
(b) $pK_a = 4.94, \Delta = -0.18$



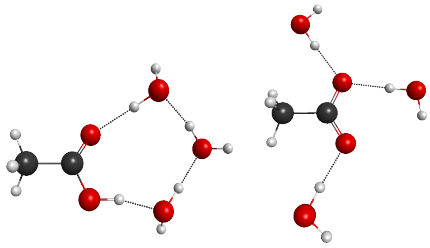
(c) $pK_a = 4.10, \Delta = 0.66$



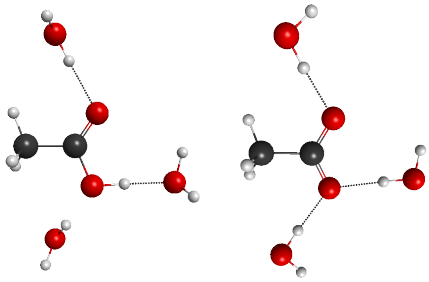
(d) $pK_a = 3.74, \Delta = 1.02$



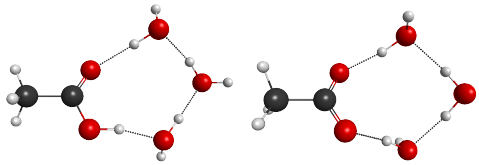
(e) $pK_a = 4.58, \Delta = 0.18$



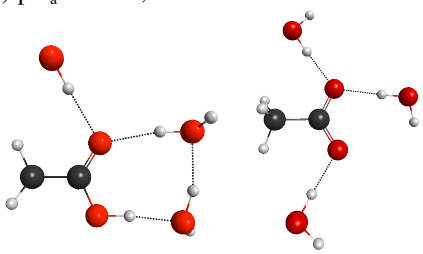
(f) $pK_a = 1.63, \Delta = 3.13$



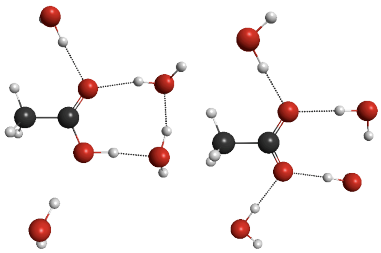
(g) $pK_a = 5.41, \Delta = -0.65$



(h) $pK_a = 4.39, \Delta = 0.37$



(i) $pK_a = 2.34, \Delta = 2.42$



(j) $pK_a = 3.55, \Delta = 1.21$

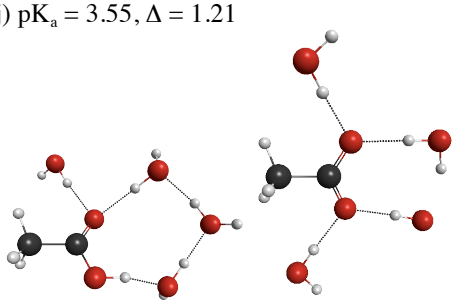


Figure 3.3.4.1.1 B97-D/6-311+G(2d,p) DSES-CC-COSab pK_a as a function of Solvation degree (S_D) and solvation sites (Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, and Q_{br}) for acetic acid and associated anion (a) $S_D(II)$, $S_C^*[Q_{a1}Q_{a2}]:S_C[Q_{a1}Q_{a2}]$; (b) $S_D(II)$, $S_C^*[Q_{a1}Q_{a2}]:S_C^*[Q_{a1}Q_{e2}]$; (c) $S_D(II)$, $S_C[Q_{a2}Q_{e1}]:S_C[Q_{a1}Q_{a2}]$; (d) $S_D(II)$, $S_C[Q_{a2}Q_{e1}]:S_C^*[Q_{a1}Q_{e2}]$; (e) $S_D(III)$, $S_C^*[Q_{a1}Q_{br}Q_{a2}]:S_C^*[Q_{a1}Q_{e1}Q_{e2}]$; (f) $S_D(III)$, $S_C[Q_{a1}Q_{e1}Q_{e2}]:S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, (g) $S_D(III)$ $S_C^*[Q_{a1}Q_{br}Q_{a2}]:S_C[Q_{a1}Q_{br}Q_{a2}]$, (h) $S_D(III)$ $S_C[Q_{a1}Q_{a2}Q_{e1}]:S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, (i) $S_D(IV)$, $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]:S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$, (j) $S_D(IV)$ $S_C^*[Q_{a1}Q_{br}Q_{a2}Q_{e2}]:S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$

3.3.4.2 Training Set: (2) Formic Acid.

Table 3.3.4.2-1 summarizes various possible sector model solvation networks, S_D, for formic acid. The basic trend in results are very similar to those found for our other training molecule, acetic acid, with the exception that the predicted pK_a results are a bit closer to the experimental value, and always underestimated. Even with no additional explicit solvent, $S_D(0)$, the predicted value is underestimated by 0.88 pK units, just outside our target of 0.74 pK units. Addition of a single explicit solvent significantly improves the predicted pK_a, with an underestimation of only 0.03 pK units. As was true for acetic acid, $S_D(II)$ for formic acid shows a preferred S_C network pair for HA:A as $S_C^*[Q_{a1}Q_{a2}]:S_C[Q_{a1}Q_{e2}]$ pair with different network for neutral and anion, but where the anion S_C is not the minimal energy configurations as it was in acetic acid. This S_D provides an underestimation in pK_a by only 0.18 pK units. The HA:A minimum energy pair S_C, $S_C^*[Q_{a1}Q_{a2}]:S_C^*[Q_{a1}Q_{a2}]$, in this case however also provides a predicted pK_a within 0.29 pK units of experiment. The remaining network combination shown is the combination that is the least preferred for the $S_D(II)$ considered (similar to acetic acid), $S_C[Q_{a2}Q_{e1}]:S_C[Q_{a1}Q_{e2}]$, which results in an underestimation of pK_a by 1.43 pK units, well outside our target of prediction.

A third explicit solvent provides two possible network combinations within our target of prediction, with a preferred $S_D(III)$ given by $S_C^*[Q_{a1}Q_{br}Q_{a2}]: S_C^*[Q_{a1}Q_{e1}Q_{e2}]$ as was found for acetic acid. $S_D(IV)$ again significantly throws off the predicted pK_a event more so than found for acetic acid, with considered network combinations being underestimated by 1.82 and 3.19 pK units.

Across all possible degrees of solvation for formic acid, the observation would indicate that a solvation degree of $S_D(II)$ with $S_C[Q_{a1}Q_{a2}]:S_C[Q_{a1}Q_{a2}]$ or $S_D(III)$ with $S_C[Q_{a1}Q_{br}Q_{a2}]:S_C[Q_{a1}Q_{e1}Q_{e2}]$ would be reasonable to considered for a 'preferred' transferable network. However, in the case of $S_D(II)$, the thermodynamically preferred $HA:A^-$ S_C pair is not the same as that predicted for acetic acid since the S_C for the anion is different ($[Q_{a1}Q_{e2}]$ in the case of acetic acid and $[Q_{a1}Q_{a2}]$ for formic acid). Although the performance of the two sets of conformations of $S_D(II)$ are not vastly different, the fact that there is a difference implies that $S_D(II)$ would not be a robust

and transferable solvent network. In contrast, the robustness and transferability of S_D(III) with only modest increase in computational complexity, make it a reasonable degree of solvation to consider as a general 'preferred' network for pK_a prediction within the tolerance set out by our model. As such, S_D(III) with $S_C^*[Q_{a1}Q_{br}Q_{a2}]$: $S_C^*[Q_{a1}Q_{e1}Q_{e2}]$ will be used to illustrate the predictability of the sector model for three variant predictive sets of carboxylic acids.

Table 3.3.4.2-1 B97D/6-311+G(2d,p) direct-sector explicit solvent in continuum model results for formic acid (exptl pK_a = 3.77)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
S_D(0)				
$S_C(0)$		3.94	2.89	+0.88
S_D(I)				
HA	**A^-**			
$S_C[Q_{a2}]$	$S_C[Q_{a2}]$	5.10	3.74	+0.03
S_D(II)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{a2}]$	$S_C[Q_{a1}Q_{e2}]$	4.89	3.59	+0.18
$S_C^*[Q_{a1}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{a2}]$**	**4.74**	**3.48**	**+0.29**
$S_C[Q_{a2}Q_{e1}]$	$S_C[Q_{a1}Q_{e2}]$	3.19	2.34	+1.43
S_D(III)				
HA	**A^-**			
$S_C[Q_{a1}Q_{a2}Q_{e1}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$	4.14	3.04	+0.73
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**4.21**	**3.09**	**+0.68**
S_D(IV)				
HA	**A^-**			
$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$	$S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$	0.80	0.58	+3.19
$S_C^*[Q_{a1}Q_{br}Q_{a2}Q_{e2}]$	**$S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$**	**2.66**	**1.95**	**+1.82**

3.3.4.3 Predictive Set (I): Increasing steric bulk: propanoic Acid, isobutryic Acid, and trimethylacetic Acid.

Given the results provided by the training set, our hypothesis is that one should be able to make accurate predictions of pK_a for carboxylic acids using the identified 'preferred' explicit solvent network, S_D(III) with $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$. While the full series illustrating the effect of additional bulk on the carboxylic moiety includes the two molecules in the training set, the actual molecules in this predictive

set are propanoic, isobutryic, and trimethylacetic acid. Tables 3.3.4.3-1 – 3.3.4.3-3 summarize the defined sector model solvation networks results in particular for the preferred network, but also including $S_D(0)$ for reference, and in the case of propanoic, the additional $S_D(II)$ networks for illustration of trends compared to the training set.

Importantly, we find that the 'preferred' network, $S_D(III)$ with $S_C[Q_{a1}Q_{br}Q_{a2}]$: $S_C[Q_{a1}Q_{e1}Q_{e2}]$, established by our training set are indeed well suited to provide predicted pK_a within our target of prediction, at -0.72, -0.23, and -0.34, for propanoic, isobutryic, and trimethylacetic acids, respectively. For reference, in all cases again, $S_D(0)$ results in an overestimation of pK_a by nearly 2 pK units. In this set, the use of identical networks for both neutral and anion throws off the pK_a prediction, as illustrated in isobutryic and trimethylacetic acid $S_D(III)$ networks. One can of course find other networks that appear to work as well as our 'preferred' $S_D(III)$ network choice, as illustrated in propanoic acid, where we have additional added $S_D(III)$ with solvent configurations for $HA:A^-$ as $S_C[Q_{a1}Q_{a2}Q_{e1}]:S_C^*[Q_{a1}Q_{e1}Q_{e2}]$. While not the thermodynamically most favored conformation for both neutral and anion, this S_D combination results in a predicted pK_a that is overestimation by only 0.40 pK units. For configurations within kT of the thermodynamic favored result, it is not surprising that apparent accurate predictions of the pK can be made. This fortuitous agreement comes about from the energy levels being close to that of the thermodynamic minimum. However, such configurations are less well defined thermodynamically and issues of arbitrariness in selecting transferable configurations will ultimately be a problem for predictability in general. Solvent configurations beyond kT of the thermodynamic minimum are not expected to be highly populated and are expected to be even less reasonable for faithful representation of the ensemble.

Likewise, there are typically many $S_D(II)$ possibilities, as illustrated for propanoic acid, where 6 combinations are shown, half of which are within our target of prediction. We note that either of the 'preferred' $S_D(II)$ suggested by our training set are outside our target range with an overestimation by 1.04 and 0.99 pK units. However, given the longer chain structure, this may be anticipated and suggests a need for more optimal placement of the two explicit water particularly encompassing Q_i positions closer to the chain. In general, the solubility of larger acids decrease very rapidly

with size, as the larger or longer substituent break up the hydrogen bonds of water replacing them by much weaker solute/solvent interactions. In this system, the thermodynamically favored configuration for the anion is actually represented by $S_C^*[Q_{e1}Q_{e2}]$, and together with the familiar thermodynamic minimum for the neutral of $S_C^*[Q_{a1}Q_{a2}]$, one finds a predicted pK_a that is just outside our target range of 0.74 pK units, with an overestimation by 0.80 pK units. Moreover, this $S_D(II)$ pair, $S_C^*[Q_{a1}Q_{a2}]$:$S_C^*[Q_{e1}Q_{e2}]$, introduces a third type of 'preferred' $S_D(II)$ type, and therefore supports the assertion that $S_D(II)$ would not be a robust and transferable solvent network.

Table 3.3.4.3-1 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for propanoic acid (exptl pK_a = 4.86)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
$S_D(0)$				
$S_C(0)$		9.19	6.74	-1.88
$S_D(I)$				
HA	**A^-**			
$S_C[Q_{a2}]$	$S_C[Q_{a2}]$	8.99	6.59	-1.73
$S_D(II)$				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{a2}]$	$S_C[Q_{a1}Q_{e2}]$	8.04	5.90	-1.04
$S_C^*[Q_{a1}Q_{a2}]$	**$S_C^*[Q_{e1}Q_{e2}]$**	**7.72**	**5.66**	**-0.80**
$S_C^*[Q_{a1}Q_{a2}]$	$S_C[Q_{a1}Q_{a2}]$	7.98	5.85	-0.99
$S_C[Q_{a2}Q_{e1}]$	$S_C^*[Q_{e1}Q_{e2}]$	6.38	4.68	+0.18
$S_C[Q_{a2}Q_{e1}]$	$S_C[Q_{a1}Q_{e2}]$	6.71	4.92	-0.06
$S_C[Q_{a2}Q_{e1}]$	$S_C[Q_{a1}Q_{a2}]$	6.64	4.87	-0.01
$S_D(III)$				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**7.61**	**5.58**	**-0.72**
$S_C[Q_{a1}Q_{a2}Q_{e1}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$	7.17	5.26	-0.40

Table 3.3.4.3-2 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for isobutyric acid (exptl pK_a = 4.88)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
$S_D(0)$				
$S_C(0)$		8.98	6.59	-1.71
$S_D(III)$				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**7.12**	**5.11**	**-0.23**
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}]$	8.09	5.93	-1.05

Table 3.3.4.3-3 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for trimethylacetic (pivalic) acid (exptl $pK_a = 5.03$)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
S_D(0)				
$S_C(0)$		9.61	7.05	-2.02
S_D(III)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{Br}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**7.33**	**5.37**	**-0.34**
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}]$	8.25	5.83	-0.80

3.3.4.4 Predictive Set (II): Electronic Substituents, Chloroacetic Acid and Glycolic Acid.

The carboxylic acid group is a highly polar organic functional group resulting from the strongly polarized carbonyl group and the hydroxyl group (Figure 3.3.4.4.1). The hydroxyl group is even more strongly polarized than in alcohols due to the presence of the carbonyl group, and together these structural features both enhance the dipole strength and impart the strong acidity of carboxylic acid compounds. Variations in R group alter the dipolar nature of these acids, and therefore the availability of Q_{a1}, Q_{a2}, Q_{e1} and Q_{e2} for energetically favorable hydrogen bonding interactions with solvent. As such, evaluation of inductive and resonance attributes of a substituent placed onto the carboxyl aid determination of number and placement of explicit water molecules required to satisfy directed interactions of the solvent network for the solute as a whole.

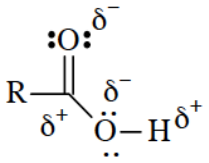


Figure 3.3.4.4.1 Depiction of the partial charge on substituted carboxyl moiety.

Consider the examples of acetic acid from the training set vs. chloroacetic acid. Chlorine as a strong electron withdrawing group (EWG) helps stabilize the negative

charge of the conjugate based formed upon ionization of the acid by electron withdrawal through carbon-carbon bonds, leaving the lone pairs on the COO(H) less available for H-bonding to water. Chloroacetic acid relative to acetic acid therefore has a substantially higher acidity, as indicated by pK_a value, 2.81 vs 4.76, respectively.

As calculated by the DSES-CC method, the pK_a of chloroacetic acid can actually be determined within our range already by $S_D(I)$, with a pK_a within 0.68 or 0.35 pK units of experiment, depending on whether the placement of the single explicit water on the primary positions of the carboxyl is syn or anti to the chloride substituent. The anti conformation is the lowest energy conformation in solution across all degrees of solvation except for $S_D(0)$. Due to the very small barrier of rotation in solvent, syn conformations can be preferred over anti conformations. Here, we discuss only the anti conformations for S_D greater than $S_D(0)$, however, results for syn are also presented in Table 3.3.4.4-1.

Addition of a second explicit water, $S_D(II)$, again gives several possibilities, a number of which are within our target of prediction, in particular, conformations suggested by our training set, $S_C[Q_{a1}Q_{a2}]$:$S_C[Q_{a1}Q_{e2}]$ and $S_C[Q_{a1}Q_{a2}]$:$S_C[Q_{a1}Q_{e2}]$ where the neutral is in the anti conformation with respect to the chloride. In this case, the network involving interaction with the two primary sites, Q_{a1} and $Q_{a2,}$ for neutral *(anti)* and anion are the thermodynamically favored conformations for both, and together provide an acceptable pK_a result, with an underestimation of 0.41 pK units. The network, $S_C^*[Q_{a1}Q_{a2}]$:$S_C[Q_{a1}Q_{e2}]$, where only the neutral is in its minimum energy conformation, provides a pK_a that is underestimated by only 0.20 pK units. Relatively poor results are obtained with networks involving $S_C[Q_{a2}Q_{e1}]$ for the neutral together with either $S_C^*[Q_{a1}Q_{a2}]$ or $S_C[Q_{a2}Q_{e1}]$ for the anion.

A third water on the network suggests again that the 'preferred' transferable network suggested by the training set, $S_D(III)$ with $S_C^*[Q_{a1}Q_{br}Q_{a2}]$: $S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, provides pK_a results within our target of prediction (+0.51) and involves the thermodynamically favored configurations for both the neutral and anion. The network $S_C^*[Q_{a1}Q_{br}Q_{a2}]$: $S_C[Q_{a1}Q_{e1}Q_{e2}]$, with a slightly different conformation of the anion but not the thermodynamically favored, provides a pK_a within 0.17 units of the experimental

value. Networks with $S_C[Q_{a1}Q_{a2}Q_{e1}]$ for the neutral in all cases result in pK_a well outside our target of prediction. Finally, one can investigate S_D(IV) possibilities, but again, as found in all other cases, the results are well outside our target of prediction and also severely underestimated.

Table 3.3.4.4-1 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for chloroacetic acid (exptl pK_a = 2.81)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
S_D(0)				
S_C(0) *anti*		2.17	1.59	+1.22
$S_C(0)^*$' *syn*		2.48	1.82	+0.99
S_D(I)				
HA	**A^-**			
$S_C^*[Q_{a2}]$ *anti*	**$S_C^*[Q_{a1}]$**	**3.35**	**2.46**	**+0.35**
$S_C[Q_{a2}]$' *syn*	$S_C^*[Q_{a1}]$	2.90	2.13	+0.68
S_D(II)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{a2}]$ *anti*	$S_C[Q_{a1}Q_{e2}]$	3.56	2.61	+0.20
$S_C[Q_{a1}Q_{a2}]$' *syn*	$S_C[Q_{a1}Q_{e2}]$	3.04	2.23	+0.58
$S_C^*[Q_{a1}Q_{a2}]$ *anti*	**$S_C^*[Q_{a1}Q_{a2}]$**	**3.27**	**2.40**	**+0.41**
$S_C[Q_{a1}Q_{a2}]$' *syn*	$S_C^*[Q_{a1}Q_{a2}]$	2.75	2.02	+0.79
$S_C^*[Q_{a1}Q_{a2}]$ *anti*	$S_C[Q_{a2}Q_{e1}]$	3.30	2.42	+0.39
$S_C[Q_{a2}Q_{e1}]$ *anti*	$S_C[Q_{a2}Q_{e1}]$	1.14	0.85	+1.96
$S_C[Q_{a2}Q_{e1}]$' *syn*	$S_C^*[Q_{a1}Q_{a2}]$	1.38	1.01	+1.80
$S_C[Q_{a2}Q_{e1}]$ *anti*	$S_C^*[Q_{a1}Q_{a2}]$	1.13	0.83	+1.98
S_D(III)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$ *anti*	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**3.07**	**2.30**	**+0.51**
$S_C[Q_{a1}Q_{a2}Q_{e1}]$' *syn*	$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$	1.80	1.32	+1.49
$S_C[Q_{a1}Q_{a2}Q_{e1}]$ *anti*	$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$	2.42	1.77	+1.04
S_D(IV)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}Q_{e1}]$	**$S_C^*[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$**	**1.81**	**1.33**	**+1.48**
$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$	$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$	-1.57	-1.15	+3.96

Glycolic acid (Table 3.3.4.4-2) provides another very clear example of how understanding the EWG/EDG inductive and resonance effects are important in deciding water placement. Figure 3.3.4.4.2(d) and (d) show S_D(II) networks, $S_C[Q_{a2}Q_{e1}]$:$S_C^*[Q_{a1}Q_{e2}]$ and $S_C[Q_{a2}Q_{e1}]$:$S_C[Q_{e1}Q_{e2}]$, respectively, where in one or both neutral and anion, explicit solvent is placed at Q_{e1}, where it can also coordinate to the hydrogen of the hydroxyl substituent on the adjacent carbon. In this case, the interaction with the explicit water molecule disrupts the normal capacity of the

hydroxyl substituent as a strong EWG, rendering these types of S_D(II) insufficient to balance the carboxylic groups directed interactions, resulting in pK_a results well outside our target of prediction. It is only for the neutral species that the Q_{el} position forms a hydrogen bond to the hydroxyl substituent. The anion maintains the hydrogen bond with the carboxyl group as this is a priority position for a directed interaction. S_D(II) conformations identified in our training set are found in this case to provide acceptable $pK_{a,}$ with $S_C^*[Q_{a1}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e2}]$ resulting in a predicted pK_a within 0.54 of the experimental value. However, addition of a third explicit water molecule around the carboxylic frame produces a sufficient solvent network for the system, where all considered S_D(III) pairs produce pK_a well within our target range, in particular, the 'preferred' conformation suggested by our training set, which provides a pK_a within 0.19 of the experimental value.

Table 3.3.4.4-2 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for glycolic acid (exptl pK_a = 3.84)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
S_D(0)				
S_C(0)		3.60	2.64	+1.20
S_D(I)				
HA	**A^-**			
$S_C^*[Q_{a2}]$	**$S_C^*[Q_{e2}]$**	**3.80**	**2.79**	**+1.05**
$S_C^*[Q_{a2}]$	$S_C[Q_{a2}]$	4.77	3.50	+0.34
S_D(II)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{a2}]$	$S_C[Q_{e1}Q_{e2}]$	5.49	4.03	-0.19
$S_C^*[Q_{a1}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e2}]$**	**4.50**	**3.30**	**+0.54**
$S_C[Q_{a2}Q_{e1}]$	$S_C^*[Q_{a1}Q_{e2}]$	2.64	1.94	+1.90
$S_C[Q_{a2}Q_{e1}]$	$S_C[Q_{e1}Q_{e2}]$	3.64	2.67	+1.17
S_D(III)				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{e1}Q_{e2}]$	4.98	3.65	+0.19
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	**$S_C^*[Q_{a2}Q_{e1}Q_{e2}]$**	**4.98**	**3.65**	**+0.19**
$S_C[Q_{a1}Q_{a2}Q_{e1}]$	$S_C[Q_{a1}Q_{e1}Q_{e2}]$	4.53	3.32	+0.52

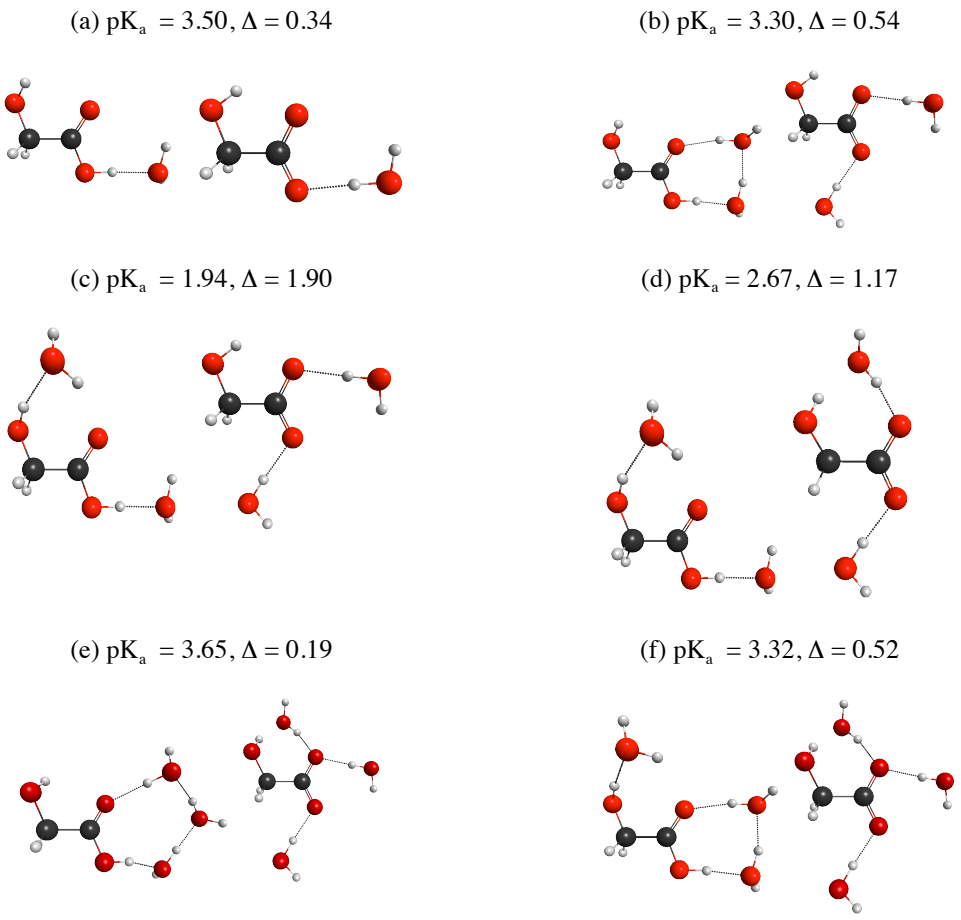


Figure 3.3.2.4.2 B97-D/6-311+G(2d,p) DSES-CC-COSab pK$_a$ as a function of S_D and solvation sites for both glycolic acid and glycolate. (a) S_D(I), $S_C[Q_{a2}]$:$S_C[Q_{a2}]$; (b) S_D(II), $S_C^*[Q_{a1}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e2}]$; (c) S_D(II), $S_C[Q_{a2}Q_{e1}]$:$S_C^*[Q_{a1}Q_{e2}]$; (d) S_D(II), $S_C[Q_{a2}Q_{e1}]$:$S_C[Q_{e1}Q_{e2}]$; (e) S_D(III), $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$; (f) S_D(III), $S_C^*[Q_{a1}Q_{a2}Q_{e1}]$:$S_c[Q_{a1}Q_{e1}Q_{e2}]$.

3.3.4.5 Predictive Set (II): Aromatic Acids, Benzoic Acid.

One final extension to consider involves COOH bonded to an aromatic ring. Aromatic carboxylic acids show both the acidity and other reactivity associated with carboxylic acids. Fundamentally, aromatic substituents do not stabilize carboxylate base charge through resonance effects. The charge created upon ionization is insulated from the aromatic ring by two single bonds. As such, aromatic rings, and substituents on aromatics rings will have only a modest effect on carboxylic acid acidity.

The simplest aromatic acid is benzoic acid (Table 3.3.4.5-1), with only slightly stronger acidity than acetic acid, 4.2 vs 4.76, respectively. As calculated by the

DSES-CC method, one finds that even without any explicit solvent, $S_D(0)$, an acceptable value of pK_a can be determined, with an overestimation with respect to experiment by only 0.50 pK units. Addition of a single explicit solvent, $S_D(I)$, disrupts the natural electronic structure of benzoic acid, and places the pK_a outside our target of prediction. The $S_D(II)$ network having the thermodynamically favored configurations of both neutral and anion, $S_C^*[Q_{a1}Q_{a2}]$:$S_C^*[Q_{a1}Q_{a2}]$, results in a pK_a of 4.47, within 0.27 of the experimental value. Notably, the $S_D(III)$ configuration suggested by our training set, $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{br}Q_{a2}]$, results in acceptable pK_a prediction of 4.70, within 0.50 of the experimental value.

Table 3.3.4.5-1 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for benzoic acid (exptl $pK_a = 4.2$)

Cluster Assignment		**ΔG**	**pK_a**	**ΔpK_a**
$S_D(0)$				
$S_C(0)$		6.41	4.70	-0.50
$S_D(I)$				
HA	**A^-**			
$S_C[Q_{a2}]$	$S_C[Q_{a2}]$	7.06	5.18	-0.98
$S_D(II)$				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{a2}]$**	**6.09**	**4.47**	**-0.27**
$S_C^*[Q_{a1}Q_{a2}]$	$S_C[Q_{e1}Q_{a2}]$	7.17	5.26	-1.06
$S_D(III)$				
HA	**A^-**			
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}]$	6.65	4.88	-0.68
$S_C^*[Q_{a1}Q_{br}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$**	**6.41**	**4.70**	**-0.50**

Table 3.3.4.5-2 categorizes all molecules considered in this study, in accord to factors discussed above and demonstrates that each class has a predictable solvent network based on the established 'preferred' network determined in the training set, using the B97-D/6-311+G(2d,p) DSES-CC-COSab methodology. In particular, the $S_D(III)$ S_C configuration pair suggested from the training set, $S_C[Q_{a1}Q_{br}Q_{a2}]$:$S_C[Q_{a1}Q_{e1}Q_{e2}]$, provides predictability within 1 kcal/mol accuracy for all systems considered here. $S_D(II)$ is more sensitive in all cases to S_C configurations, and while generally encompasses $S_C[Q_{a1}Q_{a2}]$:$S_C[Q_{a1}Q_{a2}]$, or $S_C[Q_{a1}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e2}]$ can also involve other configurations of solvation. A complete set of computed combinations of clusters with pK_a results can be found in the Electronic Supplementary Information.

Table 3.3.4.5-2 B97-D/6-311+G(2d,p) DSES-CC-COSab direct pK_a prediction using the S_C(III) 'preferred' network, as suggested by the training set.

S_D(III) Network: $S_C[Q_{a1}Q_{br}Q_{a2}]$: $S_C[Q_{a1}Q_{e1}Q_{e2}]$			
Acid	**Exptl pK_a**	**DSES-CC pK_a**	**ΔpK_a(Exptl-Calc)**
acetic acid	4.76	4.58	+0.18
formic acid	3.77	3.09	+0.68
propanoic acid	4.86	5.58	-0.72
isobutyric acid	4.88	5.11	-0.23
trimethylacetic acid	5.03	5.37	-0.34
cloroacetic acid	2.81	2.30	+0.51
glycolic acid	3.84	3.65	+0.19
benzoic acid	4.20	4.70	-0.50
		Abs. mean	0.44
		Abs. std. dev.	0.21
		Abs. max	0.72

3.3.5 Conclusions

Theoretical predictions of pK_a span a fairly large range of chemical accuracy, depending on the class of molecules considered and computational methodology chosen. While one might be able to provide an accurate pK_a prediction for a single system, accurate prediction of pK_a across an entire set of molecules using a single method is often challenging. The current work targets use of DSES-CC-QM-COSab or DSES-CC-DFT-D/COSab methods for prediction of pK_a, with the goal of investing the existence of patterns associated with placement of explicit water based on the defined-sector model presented, enabling one to subsequently contemplate understanding of any remaining small non-electrostatic energy components. Through careful consideration of solvation surfaces one can find generalizations that enable reliable determination of number and conformation of explicit solvent molecule network for classes of solutes and associated functionality. In particular, we find that there exists 'preferred' network conformations that provide pK_a within a target range of prediction of 0.74 pK units or better, S_C(III), with $S_C[Q_{a1}Q_{br}Q_{a2}]$:$S_C[Q_{a1}Q_{e1}Q_{e2}]$. One additionally can address the issue of 'flexibility' (as related to different conformations) as governed by the various conformations for the same molecule that still provide an acceptably accurate pK_a within a rigorous target of prediction. The strategy is exemplified across an important series of carboxylic acids, and shown to

have predictability within 0.74 pK units of experimental value, for a very tight range of pK_a and across varying functionality.

To apply the presented methodology as a general method for establishing explicit solvation with confidence across a broader range of systems requires further work. However, given the results presented here and the high degree of predictability, we are confident that this method is extendable as a general way to move forward. Further investigations, expanding into different classes of molecular functionality on carboxylic acids, as well as illustration and extension of the model for other classes of acids, are already underway. For example, in the former category, it is of interest to pursue other conjugating substituents and extended aliphatic chain substituents, which may introduce additional criteria to our defined-sector explicit solvent in continuum model for pK_a prediction. Importantly, it is of interest to more fully investigate S_D(IV) and S_D(V) in terms of their convergence and transferability, particularly with more extended structures, as well as degrees of solvation that further include explicit solvent at various substituent locations, Q_s. Both areas of our future work are anticipated to move the model forward for more general use, as well as enable insights into how the computational approach might be enhanced to automate the model.

3.4 Correction regarding S_D(II) conformations

In the DSES-CC analysis of the training sets, acetic acid and formic acid, the $S_C[Q_{e1}Q_{e2}]$ (Figure 3.4.1) configuration for the anions were accidentally omitted. When computed later, it was found to be the lowest energy S_D(II) configuration for both systems. Whilst this changes the observation that a consistent S_D(II) configuration pair could not be found for both acetic and formic acid, because indeed this would be $S_C[Q_{a1}Q_{a2}]$:$S_C[Q_{e1}Q_{e2}]$, tests on further systems found that it was not absolutely transferable in the way that the S_D(III) is for all the systems studied, preserving the integrity of the study and the conclusions (additional information available in Appendix A).

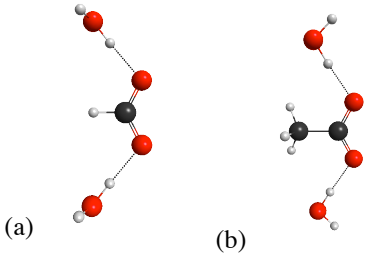


Figure 3.4.1 Depiction of the lowest energy S_D(II) configurations of (a) formate, $S_C[Q_{e1}Q_{e2}]$; (b) acetate, $S_C[Q_{e1}Q_{e2}]$

3.5 Conformational Averaging

Several authors advocate the consideration of multiple conformers in specific cases.[46, 68] In general however, most pK_a determination schemes rely on the minimum energy structure in the calculation of the free energy of the deprotonation reaction.

Indisputably, one needs to be confident that they have found the lowest energy conformation of a solute or lowest energy configuration of explicit water molecules around a solute for a given degree of solvation, S_D. Chemical intuition can largely guide choice of conformation. The defined-sector explicit solvent in continuum model approach was essentially built upon a chemical intuition regarding how the lone pairs of carboxylic/carboxylate groups interact as hydrogen donors or acceptors. However it was refined in a way that a systematic framework emerged that thoroughly considers all configurations of explicit solvent. Of less importance is the topic of conformational averaging. To exemplify this we looked at two conformers of chloroacetic acid, which is one of the acids highlighted in the paper. In conformer (1) the chloro and hydroxyl groups are anti to each other, whereas in conformer (2) they are eclipsed. Of interest with this system is that the lower energy conformer changes depending on the degree of solvation (S_D). Table 3.5-2 shows the calculated pK_a with each conformer from S_D = 0 – 3, the difference in energy between these two conformers, and also the associated anion that was used in the calculation of pK_a.

When there are multiple conformers, pK_a can be calculated by,

$$pK_a = pK_a^1 - log\chi_1 = pK_a^2 - \log\chi_2 \quad (3.5\text{-}1)$$

where X_i is the relative population of the conformer given by,

$$\chi_i = \frac{e^{-\frac{E_{rel}}{RT}}}{\sum_i^n e^{-\frac{E_{rel}}{RT}}} \quad (3.5\text{-}2)$$

where E_{rel} is the difference in energy between the lowest energy conformation and the i-conformer. R is the molar ideal gas constant equal to 1.99×10^{-3} $kcalK^{-1}mol^{-1}$ and T is the temperature in K.[69] We see that for $S_D(1-3)$, pK_a from conformational averaging is only around 0.2 pK units different from pK_a of the lower energy species (Table 3.5-1).

Whilst, conformational averaging is unlikely to be critical in regard to the possible minima of the bare solute as shown for chloroacetic acid, Chipman has raised concerns in regard to statistical sampling over the cluster configurations.[70]

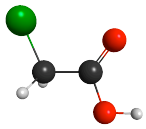
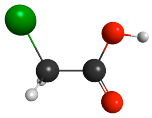
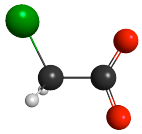
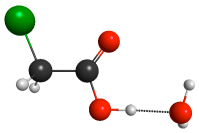
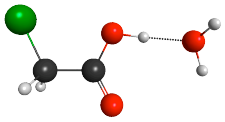
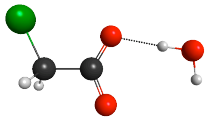
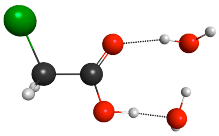
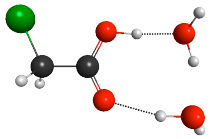
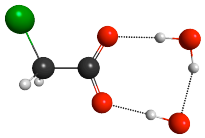
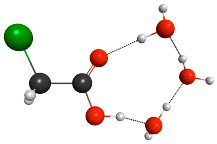
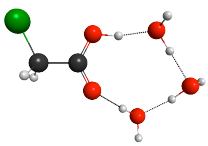
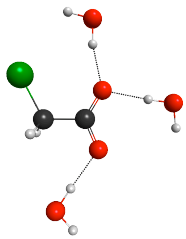
In regard to the DSES-CC configurations, only the S_C's within kT of the lowest energy conformation would need to be included, and even then, the effect is small. We took a small hypothetical experiment to get an indication of how much difference conformational averaging could change the calculated pK_a.

In this hypothetical experiment, we took the situation of there being two important conformers, such as the $S_C(Q_{a1}Q_{a2}Q_{br})$ and the $S_C(Q_{a1}Q_{a2}Q_{e1})$ configurations. Table 3.5-3 shows how the various factors in RT between the two conformations can effect the overall pK_a. If they differ by even as little as ¼ RT (0.148 kcal/mol), the change to the overall pK_a is still only 0.25 pK units, demonstrating that conformational averaging should not play an important role for the number of conformers considered in the DSES-CC model.

Table 3.5-1 Conformational averaging of chloroacetic acid

S_D	N_i/N_{total} (i = 1)	N_i/N_{total} (i = 2)	pK_a
0	0.37	0.63	2.02
1	0.68	0.32	2.62
2	0.71	0.29	2.55
3	0.69	0.31	2.41

Table 3.3.4.5-2 pK_a calculated from the two species (1 & 2), with anion (as shown on RHS) and difference in energy between the two conformers in kcal/mol

S_D	i=1 anti	i=2 eclipsed	ΔE(2-1) Kcal/mol	Lowest energy anion used in pK_a prediction
0	 pK_a = 1.59	 pK_a = 1.82	-0.32	
1	 pK_a = 2.46	 pK_a = 2.13	0.45	
2	 pK_a = 2.40	 pK_a = 2.02	0.52	
3	 pK_a = 2.25	 pK_a = 1.90	0.47	

However, low lying energy configurations arising from the rotation of the water molecules in the defined sector locations (Q_{a1}, Q_{a1}, Q_{e1}, Q_{e2}, Q_{br}) are very likely. Klamt's COSMO-RS strategy for including water molecules provides the most rigorous approach to this problem.[45a] The algorithm rolls a water molecule around the cavity of the solute and locates the point of greatest interaction.[45a] Then, it does the same to determine the directionality of the water molecule at that point on the solute surface.[45a] Examining the cluster structures from Klamt and co-workers study (available in the supplementary information)[45a], it was observed that the for many of the systems with two water molecules, the second water molecule would hydrogen bond with the first water molecule, rather than with the solute itself (Figure 3.5.1). It was reasoned that this is an artifact of adding the water molecules sequentially, as the point of highest charge may be transferred to the first water molecule, but it is unlikely that the solute would only have one water molecule in its primary solvation shell and thus an unintuitive cluster emerges. This problem currently prohibits Klamt's methodology from being exploited, however, further development of the DSES-CC model should consider the directionality of the explicit water molecules. A number of the issues relating to statistical sampling and structure characterization will be addressed in the Chapter 5.

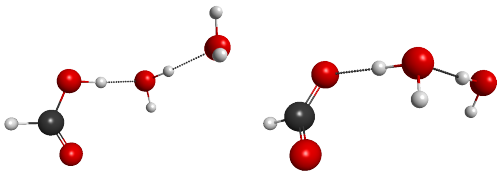


Figure 3.5.1 Klamt and co-workers S_D(II) conformations for formic acid and formate

Table 3.5-3 Change to the pK_a of the lower energy species when a second conformer is taken into consideration at a factor X of RT from the lower energy species.

Factor of RT	Energy kcal/mol	Change to the pK_a of the lower energy species
0.25	0.148	2.50E-01
0.5	0.296	2.06E-01
0.75	0.444	1.68E-01
1	0.592	1.36E-01
2	1.184	5.51E-02
3	1.776	2.11E-02
5	2.96	2.92E-03
10	5.92	1.97E-05

4 Applying the DSES-CC model

4.1 Introduction

The DSES-CC methodology established in the previous chapter provides a transferable approach to adding explicit water molecules in the cluster-continuum methodology for carboxylic acids. Whilst the first publication, "Defined-Sector Explicit Solvent in Continuum Model Approach for Computational Prediction of pK_a,"[35] looked at some interesting functionality including systems with increasing bulk, electron withdrawing groups and aromatic acids, it was of interest to see how the methodology behaved for a larger range of carboxylic acid systems, including systems that have secondary functionality in the substituent shell. This chapter, which includes the second publication using the DSES-CC, model extends the study to include:

1) An expansion of predictive set A; increasing bulk, electron withdrawing substituent groups, and unsaturated systems.
2) Expansion of predictive set B, aromatic acids, to include secondary functional groups.
3) A new predictive set, C, dicarboxylic acids.
4) The second dissociation constants of the dicarboxylic acids.

4.2 Defined-Sector Explicit Solvent in Continuum Cluster Model for computational prediction of pK_a: Consideration of secondary functionality and higher degree of solvation.

Authors: Rebecca A. Abramson and Kim K. Baldridge

This work is published in J. Chem. Theory Comput. (2013) 9 pp. 1027 – 1035.

4.2.1 Introduction

Accurate prediction of acid dissociation constants (K_a) has seen significant progress in recent literature.[60-61, 71] A first principles prediction of pK_a within 0.5 pK units of experimental values has been a challenge for the theory of proton transfer reactions, and has therefore become a benchmark of broad interest.[28, 45a, 45c, 45d, 46, 50] The inherent challenge for QM methods is that at ambient temperature as little as 0.7 kcal/mol error in the ΔG_{diss} leads to misestimation of the pK_a by the benchmark 0.5 pK_a unit, whereas +/- 1.0 kcal/mol accuracy in energy is still a difficult level to achieve using *ab initio* solvent strategies. To address this challenge, we recently developed the defined-sector explicit solvent in continuum model (DSES-CC) approach, which enables a systematic approach for predictability of solvent networks based on an established preferred conformation of explicit solvent to within +/- 1.0 kcal/mol.[35] The model was demonstrated through consideration of the structure-to-chemical affinity relationship of the carboxyl functional group.[51]

The defined-sector model provides a systematic basis for inclusion of explicit water molecules in the molecular cavity embedded in implicit solvent, as the continuum-cluster (or explicit-implicit) method. In this method, pK_a is calculated directly from the continuum-cluster method, without using a thermodynamic cycle or means of fitting to experiment. Clusters are systematized based on a strategy for placement of the explicit solvent molecules with respect to the solute. Specific solvation states are defined according to degree of solvation (S_D) and configuration of solvation (S_C). The degree of solvation (S_D) is defined as the number of explicit solvent molecules needed, and the configuration of solvation (S_C) is defined by the specific set of principle solvation sites, secondary solvation sites, and sites within the substituent shell, where

solvent is explicitly placed. For the particular case of carboxylic acid and carboxylate functionality, the principal and secondary explicit solvation sites can be illustrated as in Figure 4.2.1.1. Depending on the nature of the substituents on the carboxylic acid, the substituent shell will accommodate explicit solvation, $S_D(N + M)$, where N refers to the degree of solvation of the primary carboxylic moiety and M refers to the degree of solvation of the substituent shell. Evaluation across an array of S_D's reveals patterns of limited direct solvation, and provides an indication of how various S_C's affect prediction of pK_a for a set of molecules.

In the present work, additional development of the DSES-CC model for prediction of pK_a is illustrated across a much broader set of carboxylic acids (>32, including 9 dicarboxylic acids), thereby further substantiating the model for general use. A much broader range of electronic structure functionality is now addressed, including issues of substituent shell explicit solvation. Important to the fundamentals of the continuum model approach in general, higher degrees of solvation are explored up to $S_D(V)$, which fills all degrees of solvation for the carboxylic acid functionality (Figure 4.2.1.1). Finally, prediction of pK_a for dicarboxylic acids including prediction of the pK_a^2 is undertaken with the DSES-CC model.

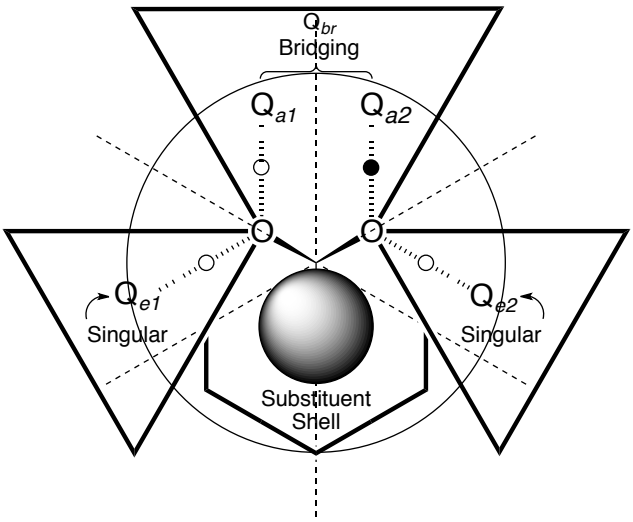


Figure 4.2.1.1 Depiction of principal and secondary explicit solvation sites around a carboxylic acid (or carboxylate). Small circles indicate presence (filled) or absence (open) of H.[35]

4.2.2 Computational Methods

All calculations were performed with the GAMESS electronic structure program.[4] Full optimizations were carried out including effects of solvation via the DSES-CC model, using B97-D/6-311+G(2d,p)/COSab, with our most recent implementation of COSab solvation model.[3, 33, 56] Parameter optimization for several combinations of DFT functional type and basis sets have been carried out within the solvation model in previous work.[37b] In our initial development of the DSES-CC model, investigation covering basis set, wave function type, thermodynamic cycle, reaction scheme, and solvent parameters, was carried out.[35] As with any property, one should expect to find variation with DFT-type, basis set representation, and solvent specifications, so it is important is to choose a functional that is appropriate for the property.[28, 72] In particular, methodology should accommodate the weak interactions present in the explicit/implicit solvent systems. The present work as well as our previous studies well supports the reliability of the B97-D functional together with a triple-z basis set. The dispersion enabled density functional B97-D is a reparameterization of the original B97 hybrid functional of Becke,[53] and has been implemented and tested in GAMESS within the solvent model. [37b] An ultrafine grid, NRAD=96 NLEB=1202 was specified. The triple-z basis set representation 6-311+G(2d,p),[73] was employed. Analytic hessian calculations were carried out to characterize the structures and determine zero point energy corrections. Dielectric permittivity of water (ε=78.4) was used, with cavity parameters of 1082 points for the basic grid, 92 segments on the complete sphere. Outlying charge error correction was taken into account via the double cavity approach.[3] DSES-CC representations were depicted using MacMolPlt.[59]

Consideration of contributions to the non-electrostatic solvation term, most importantly cavitation and dispersion-repulsion, is important for calculation of accurate pK_a. Under the assumption that the differential cavitation term between carboxylic acids and carboxylates is negligible, inclusion of directed effects through explicit consideration of primary waters of hydration should enable a high level of accuracy if explicit solvation is properly handled. The defined-sector explicit solvent in continuum cluster (DSES-CC) model relies only on solution phase computations (i.e., eliminating the use of a thermodynamic cycle or fitting schemes) together with

the sector model for placement of explicit solvent molecules. This method eliminates a number of possible sources of error, making use of the reaction scheme,

$$AH.(H_2O)_n \rightarrow A^-.(H_2O)_n + H^+$$

Both experimental and theoretical values have been used for the free energy of the proton in the literature, due to the associated difficulties for determining this quantity directly.[66] We agree with the previous thorough investigations in the use of the value -265.9 kcal/mol.[44a, 45d, 50a, 67] The gas phase energy is indisputably derived from an enthalpy contribution, 2.5RT, and an entropic contribution calculated from the Sackur-Tetrode equation, yielding a value of -6.28 kcal/mol.[63c] Unique to the DSES-CC model is a greater depth of analysis involving networks of explicit solvent molecules on the individual components of the proton transfer reaction. The solvation state energy is determined for each component of the acid dissociation reaction ($HA•S_C[Q...]$ or $A^-•S_C[Q...]$) in a specific S_C within a given S_D. The ΔG of any specific S_C is determined by subtracting the energy of the reactant state from the product state ($\Delta G = (A^-• S_C[Q...] + H^+) - HA• S_C[Q...]$). The lowest energy set of S_C within a given S_D (labeled S_C^*) is used to determine the thermodynamic ΔG_{diss} of acid dissociation for a given S_D. The pK_a follows directly as $\Delta G/2.3RT$,[52] and the calculated value is compared to the experimental value as $\Delta pK = pK_a(expt)-pK_a(calcd)$.

4.2.3 Results

4.2.3.1 Initial Predictive Set.

In our first study, a set of carboxylic acids spanning several classes of functionality was investigated.[35] A training set was used to identify a thermodynamically transferable preferred solvent network, which was then applied to three categories of acid structure functionality. Evaluation criteria was based on the fact that ½ a pK unit is, in energy terms, only 0.68 kcal/mol, so an acceptable range of predictability was defined to be within 1 kcal/mol of the experimental value, or, 0.74 pK units. Within the DSES-CC model, it is possible that a range of 'acceptable' HA/A^- pairs for a given $S_D(X)$ may provide pK_a prediction within this target range; however, among any

range of potentially acceptable S_C pairs, only S_C within kT of the thermodynamically favored pair need be considered, as others would not be energetically feasible.

Categorization based on electronic and resonance substituents can provide rationalizations for the best S_D and S_C's for each of the different predictive groupings; however, the ultimate goal was to provide a robust transferable cluster that provides consistent results across a large set of compounds within the target range of 0.74 pK units (+/-1 kcal/mol) of the experimental value. The initial findings showed that S_D(I) clusters generally fail and were only found to be sufficient for electron withdrawing substituents. Although S_D(II) configurations can produce accurate prediction for the small set, there is not a particular S_C that serves across all systems, and consequently does not offer the desired transferability. On the other hand, a specific S_D(III) cluster, $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, enables pKa prediction within 1 kcalmol-1, or 0.74 pK units, across this entire set of carboxylic acids, as a transferable S_C (Table 4.2.3.2-1, first 8 acids). Figure 4.2.3.1.1 shows an example of the S_D(III) for one of the training set of acids, acetic acid. These initial studies demonstrate that, through careful consideration of solvation networks, one can assess their predictive power as a function of the number and conformation of explicit solvent molecules, for specific classes of solutes. Such a systematic assessment offers the chance to extend the explicit solvation model to establish general methods applicable to a broader range of solutes.

To establish the proposed methodology as a general method for the positioning of explicit solvation across a broad range of carboxylic acids, the initial set was broadened to include several other substituted carboxylic acids (Predictive Sets A), additional degrees of solvation through the substituent shell (Predictive Sets B), and higher order pK_a prediction (Predictive Sets C). In addition, discussion of higher degrees of solvation, S_D(IV) and S_D(V). Each of these is discussed in detail in what follows.

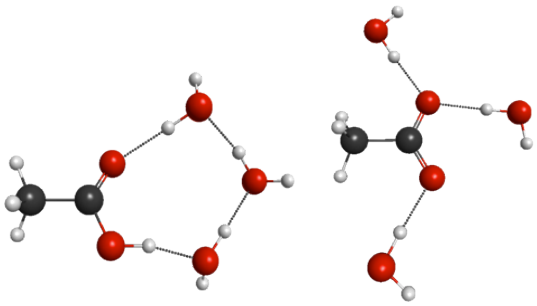


Figure 4.2.3.1.1 Depiction of S_D(III) for acetic acid HA and A^- pair. Experimental and calculated values of pK_a are 4.7_6 and 4.5_8, respectively.

4.2.3.2 Expanded Predictive Sets A.

Given the results provided by the chosen training set (acetic and formic acid), our original hypothesis was that one should be able to make accurate predictions of pK_a for any carboxylic acid using the identified 'preferred' explicit solvent network, S_D(III) with $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$. It was possible to show this to be the case for three predictive sets of carboxylic acids, including (I) a class with increasing steric bulk (electron donating groups), (II) a class with electronic withdrawing groups, and (III) an aromatic carboxylic acid functionality. In the present study, a greatly expanded set of acids has been included to probe further the predictability of the DSES-CC model, using the same level of theory as our initial study.[35] In particular, the initial trio of predictive sets has been now expanded to include extended substituent bulk (electron donating) in class I, additional electron withdrawing substituent groups in class II, a more extensive look into aromatic acids beyond the original single system, and a new predictive set of unsaturated functionality, set (IV).

In general, the expectation is that electron withdrawing/donating groups will influence the acidity of a carboxylic acid primarily through stabilization/destabilization of the conjugate base, i.e., inductive effects, resulting in an increase/decrease in the acidity of the acid. Additionally, in unsaturated analogues, delocalization of charge through resonance will be a further charge stabilizing effect, altering the acidity. It is the balance of inductive and resonance effects as partitioned in the mind of the chemist

that must be properly modeled computationally, including the important explicit solvation interactions, for accurate prediction of pK_a in these systems.[74]

The extended series of predictive set I in the initial study, illustrates the effect of additional bulk on the carboxylic moiety. The full series includes acetic, formic, propanoic, isobutyric, trimethylacetic, butanoic, pentanoic, and cyclohexane carboxylic acids. One can observe the effect of longer saturated chains on the carboxylic acid functionality within the series propanoic, butanoic, and pentanoic acids. In particular, one could imagine that the saturated tail might require additional explicit waters of solvation; however, it appears that no additional substituent shell interactions are necessary to provide pK_a within the tolerance set out. Similarly, other bulky additions to the carboxylic acid does not appear to require attention with respect to additional explicit waters around the substituent group. Data on the preferred S_D(III) network for the full predictive set I show that this network is indeed well suited to provide predicted pK_a within the target of prediction, with deviations of $+0.1_8$, $+0.6_8$, -0.7_2, -0.2_3, -0.3_4, -0.4_1, -0.4_2, and -0.7_2 pK_a units from experiment, for the above series members, respectively (Table 4.2.3.2-1). For reference, in all cases $S_D(0)$ results show an overestimation of pK_a by ~2 pK units.

The extended series of predictive set II in the initial study, illustrating the effect of electron withdrawing groups, includes chloroacetic, glycolic, nitroacetic, and mandelic, acids. The electron withdrawing groups were considered with regard to how they modify the dipolar nature of the carboxylic acid scaffold, and consequently the availability of the principal and secondary solvation sites, Q_{a1}, $Q_{a2,}$ Q_{e1}, Q_{e2}, and Q_{br}. Nitroacetic acid (experimental pK_a of 1.3_2), offers an even stronger electron-withdrawing group than chloroacetic acid, therefore testing the robust nature of the preferred solvent configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$ for carboxylic acids with very low pK_a values. In this case, pK_a prediction was only 0.1_7 pK units from the experimental value, well within the target deviation. Mandelic acid was considered as an analogue of glycolic acid. The preferred solvent configuration performs well, with the predicted pK_a value only 0.3_0 units below the experimental value of 3.4_1. The results from this predictive set of electron withdrawing substituents are important as they demonstrate that, even with very strong electron withdrawing groups that offer

significant stabilization of the carboxylate charge, the solvation sites identified by the preferred configuration, Q_{a1}, Q_{a2}, Q_{br}, (acid) and Q_{a1}, Q_{e1}, Q_{e2} (anion) suffice for accurate predictions.

Predictive set IV introduces the important class of unsaturated carboxylic acids, in particular the 'ene' functionality. The inductive effect of the 'ene' functionality serves to stabilize the carboxylate relative to the acid; however, the resonance contribution can play a role in stabilizing the carboxylic acid state. This set includes acrylic, crotonic, and cinnamic acids. (Note that the general treatment of aromatic acids is treated as a separate predictive set.) In the first two acids of the series, acrylic and crotonic, application of the preferred S_D(III) configurations predicts a pK_a only slightly below that of acetic acid, and is well within the target tolerance, with deviations from experimental values of 0.2_9 and -0.3_9 units, respectively (cf. Table 4.2.3.2-1).

A particularly difficult unsaturated carboxylic acid is trans-cinnamic acid (3-phenylacrylic acid), where there is an additional phenyl substitution on the 'ene' functionality. In this case, the preferred S_D(III) configuration results in a predicted pK_a just outside the target range ($\Delta = 0.9_6$ from experiment). The associated resonance structures of cinnamic acid suggest the need to provide explicit water interactions at the Q_{e1} and Q_{a1} positions in both HA and A^-, in addition to the single explicit water at Q_{a2} and Q_{e2}, for HA and A^-, respectively. In fact, an S_D(III) configuration of $S_C[Q_{a1}Q_{a2}Q_{e1}]$ around the acidic species (< 1 kcal/mol from thermodynamic minimum) together with the standard configuration $S_C[Q_{a1}Q_{e1}Q_{e2}]$ around the anionic species, results in a calculated pK_a (4.6_8) within 0.2_4 units of experiment. In this case, the acid is an electron deficient group that can be stabilized by resonance contributions from the alkene acting as a donor, but the carboxylate is electron rich and cannot benefit from the donor forms of the alkene. The alkene functions then as an electron withdrawing group on the carboxylate through inductive effects only, because there are no beneficial resonance forms shifting electron density from the carboxylate to the alkene. In the acid, the alkene serves as a better donor because its resonance forms are further stabilized by contributions from the phenyl ring.

The collective expanded predictive set A is shown in Table 4.2.3.2-1 (additional details in Appendix A), with 16 acids using the transferable S_D(III) configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]:S_C^*[Q_{a1}Q_{e1}Q_{e2}]$. The set shows predicted pK_a well within the target range of 0.74 pK units using only the principal and secondary sites of the carboxylic acid functionality. The single exception discussed is cinnamic acid, where the preferred S_D(III) predicts a pK_a just outside the tolerance (calcd. error 0.9_6 kcal/mol), but for which an acceptable result is found with an alternative S_D(III) where the specific configuration is based on resonance considerations. The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.44 (std. dev. 0.23).

4.2.3.3 Expanded Predictive Set B, Aromatic Acids

Predictive set B greatly expands on the class of aromatic acids, which consisted of 1 aromatic acid (benzoic acid) in predictive set III of the initial study. Analogous to the 'ene' functionality in Predictive Set A, the carboxylic acid, as an electron deficient group, can be stabilized by resonance contributions from the aromatic ring acting as a donor. The electron rich carboxylate cannot benefit from the donor forms of the aromatic ring, which instead affects the carboxylate through induction. The primary resonance forms of the acid provide insight into how the functionalized aromatic substituent affects the acidity. This will in turn provide insight into first solvation shell explicit interactions, including the need for explicit solvent representation in the substituent shell of the aromatic component.

In the initial trio of predictive sets, the aromatic functionality on the carboxylic acid was briefly investigated with the simplest aromatic acid, benzoic acid. In this case, the ring has only a small influence on the acidity of the carboxyl unit such that benzoic acid is only a slightly stronger acid than acetic acid (4.2_0 vs. 4.7_6, expt; 4.7_0 vs. 4.5_8, calc.). The relatively weak affect of the phenyl substituent is a result of the additional resonance effect in the acid that is not present in the anion. The result is a weaker acid than what might be expected. Computation predicts benzoic acid to be less acidic than experimentally observed, and acetic acid to be slightly more acidic

Table 4.2.3.3-1 B97-D/6-311+G(2d,p) DSES-CC-COSab direct pK_a prediction using the 'preferred' solvent network, S_D(III) and S_D(III+M) (M=substituent coverage), for carboxylic acids set compared to experiment.[1]

Acid	S_D	Exptl. pK_a	DSES-CC pK_a	ΔpK_a
Initial Predictive Set				
acetic	S_D(III)	4.7_6	4.5_8	0.1_8
formic	S_D(III)	3.7_7	3.0_9	0.6_8
propanoic	S_D(III)	4.8_6	5.5_8	-0.7_2
isobutyric	S_D(III)	4.8_8	5.1_1	-0.2_3
trimethylacetic	S_D(III)	5.0_3	5.3_7	-0.3_4
chloroacetic	S_D(III)	2.8_1	2.2_5	0.5_6
glycolic	S_D(III)	3.8_4	3.6_5	0.1_9
benzoic	S_D(III)	4.2_0	4.7_0	-0.5_0
Expanded Predictive Set A				
butanoic	S_D(III)	4.8_3	5.2_4	-0.4_1
pentanoic	S_D(III)	4.8_4	5.2_6	-0.4_2
cyclohexanecarboxylic	S_D(III)	4.9_0	5.6_2	-0.7_2
nitroacetic	S_D(III)	1.3_2	1.4_9	-0.1_7
mandelic	S_D(III)	3.4_1	3.1_1	0.3_0
acrylic	S_D(III)	4.2_6	4.5_5	-0.2_9
crotonic	S_D(III)	4.6_9	5.0_8	-0.3_9
trans-cinnamic	S_D(III)	4.4_4	5.4_0	-0.9_6
	S_D'(III)[2]		4.6_8	-0.2_4
Predictive Set B				
o-hydroxybenzoic	S_D(III)	2.9_8	2.3_0	0.6_8
m-hydroxybenzoic	S_D(III)	4.0_8	4.5_2	-0.4_4
	S_D(III+I)		3.9_2	0.1_6
p-hydroxybenzoic	S_D(III)	4.5_8	5.0_4	-0.4_6
	S_D(III+I)		4.4_5	0.1_3
p-methoxybenzoic	S_D(III)	4.5_0	5.3_7	-0.8_7
	S_D(III+I)		5.2_4	-0.7_4
p-butylbenzoic	S_D(III)	4.4_7	5.0_2	-0.5_5
p-aminobenzoic	S_D(III)	4.9_2	6.3_1	-1.3_9
	S_D(III+I)	4.9_2	5.8_6	-0.9_4
p-nitrobenzoic	S_D(III)	3.4_0	3.1_9	0.2_1
Predictive Set C – pK_{a1}				
carbonic[3]	S_D(III)	3.5_8	2.2_3	1.3_5
	S_D(III+I)		3.0_5	0.5_3
oxalic	S_D(III+I)	1.2_3	1.4_4	-0.2_1
malonic	S_D(III+I)	2.8_3	3.3_9	-0.5_6
succinic	S_D(III+I)	4.1_6	5.0_4	-0.8_8
	S_D(III+II)		4.9_4	-0.7_8
	S_D(III+III)		4.9_5	-0.7_9
adipic	S_D(III+I)	4.4_3	5.2_3	-0.8_0
	S_D(III+III)		5.1_8	-0.7_5
fumaric	S_D(III+I)	3.0_3	4.1_6	-1.1_3
	S_D(III+III)		3.7_8	-0.7_5
Maleic	S_D(III+1)	1.8_3	2.5_7	-0.7_4
terephthalic	S_D(III+1)	3.5_1	4.0_7	-0.5_6
cyclohexanedicarboxylic	S_D(III+1)	4.1_8	4.96	-0.7_8
Predictive Set C – pK_{a2}				
carbonic[3]	S_D(V)'	10.6_0	10.9_5	-0.3_5
oxalic	S_D(III+III)	4.1_9	4.6_2	-0.4_3
malonic	S_D(III+III)	5.7_9	5.4_8	-0.0_1
adipic	S_D(III+III)	5.4_1	5.8_5	-0.4_4
succinic	S_D(III+III)	5.6_1	5.5_1	0.1_0
fumaric	S_D(III+III)	4.4_4	4.6_6	-0.2_2
Maleic	S_D(III+III)	6.0_7	6.0_4	0.0_3

terephthalic	S_D(III+III)	4.4_0	5.1_9	-0.7_9
cyclohexanedicarboxylic	S_D(III+III)	5.4_2	6.1_7	-0.7_5

[1]Experimental values were taken from:
G. Kortum, W. Vogel, K. Andrussow, Dissociation Constants of Organic Acids in Aqueous Solution, Butterworths Scientific Publications, London, 1961
A. Klamt, F. Eckert, M. Diedenhofen, M.E. Beck, 2003, J. Phys. Chem A, vol. 107 pp. 9380-9386
[2]See text for discussion of explicit solvent for cinnamic acid.
[3]See text for discussion of explicit solvent for carbonic acid.

than observed. However, this is a result of the computational model giving greater weight to resonance effects in the carboxylic acid compared to the inductive effects in the carboxylate. As the two systems are experimentally very close in pK value, the balance of the two effects plays an important role in predicting rank order, even if the model provides good results for each independently.

Substituents on the aromatic ring further alter the acidity of carboxylic acids through inductive and/or resonance effects, depending on the nature, type, and placement on the ring. In general, one expects an increase in acidity (lower pK_a) with addition of electron withdrawing substituents on the aromatic ring, and a decrease in acidity (higher pK_a) with electron donating groups on the aromatic ring.[74] Consideration of hydroxyl-, methoxy-, amino-, butyl-, and nitro- benzoic acid derivatives enables further testing of the DSES-CC model, in terms of the transferable S_D(III) network, and illustrates the need for further explicit interactions in the substituent shell.

Investigations of the three (o-,m-,p-) isomers of hydroxybenzoic acids provide an interesting test of the DSES-CC model, as the balance of effects varies with position of substituent on the ring, resulting in significant variation in acidity of the three (exptl values 2.9_8, 4.0_8, and 4.5_8, respectively). The para- derivative is the least acidic of the three isomers relative to benzoic acid, considering only an inductive effect. In addition, the para isomer also has an important resonance effect deriving from the hydroxyl- substituent resonating into the ring and through to the carboxylic acid. This effect stabilizes the acid form, but not the anion form, resulting in the lower acidity of the system compared to the other isomers. This resonance effect is not important in the meta- derivative, and as such, the affect of the m-hydroxybenzene substituent on the carboxyl unit is primarily inductive in nature, resulting in an only slightly more acidic system than benzoic acid. Importantly, in the para and meta isomers, the preferred S_D(III) predicts a pK_a value within the tolerance limit: -0.4_4 and -0.4_6,

respectively. One might consider further the need to add a single explicit water molecule interacting with the lone pair of the hydroxyl substituent of the aromatic moiety, S_D(III+1). In this case, the model predicts a pK_a value within given tolerance limits (0.1_3 and -0.1_6 for para- and meta- respectively). Therefore, these results suggest that both S_D(III) with $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, as well as S_D(III+1) with $S_C[Q_{a1}Q_{br}Q_{a2};Q_S]$:$S_C[Q_{a1}Q_{e1}Q_{e2};Q_S]$ satisfy our criteria and offer good prediction of the pK_a value.

Although one finds a similar resonance delocalization for salicylic acid (o-hydroxybenzoic acid) and an opposing inductive effect, the proximity of the hydroxyl substituent to the carboxyl units allows for a favorable intra-molecular hydrogen bond to be present in the latter given the anion negative charge. The combined effect is a much stronger acid, with a predicted pK_a value of 2.3_0 (exptl, 2.9_8). In terms of explicit water interactions, the intra-molecular interaction serves to reduce the number of explicit water molecules interactions needed in the first solvation shell. Coincidentally, S_D(I) provides a pK_a result 0.0_3 pK_a units from the experimental value; however, the S_D(III) preferred configuration results in thermodynamically favored configurations, with predicted pK_a value within the tolerance limits 0.6_8 below the experimental result.

Modification of the hydroxy substituent in p-hydroxybenzoic acid to p-methoxybenzoic acid allows a further test of the sensitivity of the DSES-CC. Calculations with the preferred configurations around the carboxylic/carboxylate moieties results in overestimated pK_a values by 0.8_7, which is slightly outside the tolerance limit. In this case, however, the availability of the methoxy lone pair is attenuated by the inductive effect of the methyl group in comparison to the hydroxyl unit. As such, addition of an explicit water molecule in the substituent shell is warranted here, and in fact, improves the calculated pK_a value to within the tolerance limit of 0.7_4 pK units. Alteration in aromatic substituent from alkoxy to alkyl, as in para-butylbenzoic acid, results in a substituent that is inductive, and predictions using the DSES-CC model with the preferred S_D(III) configuration gives the pK_a only 0.5_5 units above the experimental value.

Replacing the aromatic substituent with an electron-withdrawing nitro group serves to increase the acidity of the carboxylic acid, as it stabilizes the parent acid. The S_D(III) preferred configuration in this case provides a prediction of pK_a value for para-nitrobenzoic acid within the tolerance limit, 0.2_1 units below the experimental value, at 3.1_9 (exptl. 3.4_0).

A more difficult case is amino-substitution, where the amino substituent is an electron-donating group through resonance and electron withdrawing through induction. In this case, it is necessary to consider substituent shell explicit solvent interactions with the lone pair of the amine group, as pK_a predictions are over 1 pK unit too acidic without consideration of explicit solvent on the amino group. Using the preferred configuration of solvation for acid and anion, plus additional solvent shell representation, the best estimate is just outside the target range at 0.9_4 pK units too basic, however, still within 1.2_0 kcal/mol of experiment, and so considered acceptable given the known difficulties in modeling amino functionality.

The collective expanded predictive set B (cf. Table 4.2.3.2-1 and Appendix A) with 7 acids using the transferable S_D(III) configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, together with 0 or 1 additional explicit solvents on the aromatic substituent depending on the nature of the aromatic substituent, shows predicted pK_a well within the tolerance limits set out. The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.58 (std. dev. 0.24).

4.2.3.4 Expanded Predictive Set C, Diprotic Acids: dicarboxylic acids.

Thus far, only carboxylic acids with a single ionizable group have been considered, and the pK_a value rationalized via the DSES-CC model with respect to the various structural features of the acid. Another important test for the DSES-CC model is the class of polyprotic acids, which have presented significant challenge for prediction of pK_a values.[71b, 75] The grouping of polyfunctional acids with general formula, C(O)OH-R-C(O)OH (R=alkyl, alkenyl, alkynyl, aryl), is characterized by having two ionizable carboxylic acid units. There are a number of issues pertaining to the

prediction of pK_a values of these acids, including (1) whether the preferred S_D(III) network provides adequate prediction for the first deprotonation reaction, given the change in electronic structure due to the presence of the second carboxylic group, (2) whether the second carboxylic group should be treated as a substituent, or, supports also the preferred S_D(III) configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, and (3) whether calculation of the pK_a value of the second deprotonation reaction using the DSES-CC method also provides predictive results. Points (1) and (2) are addressed in this section, point (3) is addressed in the following section.

To address points (1) and (2), it is instructive to consider more specifically what constitutes the DSES-CC model for such a system (cf. Figure 4.2.1.1). If the second carboxylic acid group is considered as part of the substituent shell, then the number of principal and secondary explicit solvent molecules does not change, and one only needs to address any needed substituent shell explicit solvents, in much the same way as already treated in the monocarboxylic acids. If, on the other hand, one considers each of the two carboxylic acid moieties as primary sites of explicit solvation, then the number of principal and secondary solvation sites exactly doubles, Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, and Q_{br}, and consideration of potential Q_s sites in between the carboxylic acid functionalities must also be addressed. In the latter case, the relative positioning of the two carboxylic acid functionalities with respect to one another could allow for shared Q_{e1} and Q_{e2} principal sites (e.g., oxalic acid).

For the series of alkyl dicarboxylic acids, acidity is related to the chain length of the alkyl group between the two carboxylic groups. In the series considered here, carbonic acid is included as it has been used as an exemplary case in a number of studies in the prediction of pK_{a2} value.[45d, 64d] Figure 4.2.3.4.1 gives a depiction of the principal and secondary explicit solvation sites within the DSES-CC model in this special case. For the first deprotonation reaction of carbonic acid to bicarbonate anion a selection of possible S_C's of the primary carboxylic group is shown in Figure 4.2.3.4.2 (see also Appendix A). Importantly, (h) is the preferred S_D(III+1) configuration, which provides predicted pK_a value within 0.5_3 of the experimental value. This example illustrates the flexibility of the DSES-CC model to treat a special case.

Oxalic acid, HOOCCOOH, is the shortest chain with two separate carboxylic acids. In this case, one expects the first pK_a value to be significantly lower than the typical monocarboxylic acid because formation of the mono-anion is facilitated (stabilized) by the residual acid via hydrogen bonding. Prediction of the pK_a value for oxalic acid was achieved using the preferred S_D(III) explicit solvent configuration applied to one of the carboxylic units and a single explicit water applied to the second carboxylic group, thereby treating the second carboxyl unit as a substituent. The predicted pK_a value is indeed quite acidic at 1.4_4, and the result is well within the tolerance limit of the experimental value of 1.2_3 (Table 4.2.3.2-1).

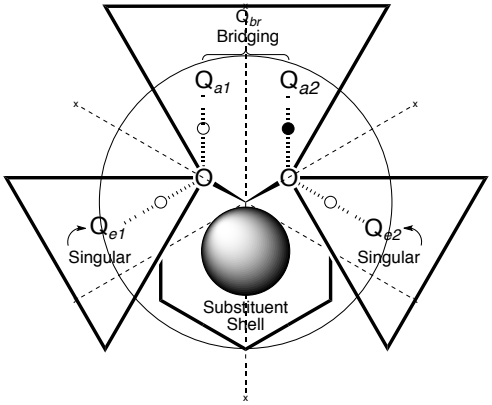


Figure 4.2.3.4.1 Depiction of the principal and secondary explicit solvation sites around dicarbonic acid (or the deprotonated forms). Small circles indicate presence (filled) or absence (open) of H.

When the number of carbon atoms between the carboxyl units increases, as in the series: malonic, $COOHCH_2COOH$, succinic, $COOHCH_2CH_2COOH$, and adipic, COOH $CH_2CH_2CH_2CH_2COOH$, acids, geometric constraints and strong local solvation from water prevent the formation of stabilizing intramolecular H-bonds, resulting in acids that are much less acidic than oxalic acid. Computations using S_D(III+1) predicted pK_{a1} values for malonic, succinic, and adipic acids of 3.3_9 (2.8_3), 5.0_4 (4.1_6), and 5.2_3 (4.4_3), respectively, where values in parenthesis are experimental results (cf. Table 4.2.3.2-1). These results suggest that the affect of the second carboxyl unit is as a substituent. In addition, the influence of applying 1, 2, or 3

explicit solvents is relatively minor, but does show a convergence of results from S_D(III+1), to S_D(III+II), to S_D(III+III), such that all values come within the tolerance limits. Table 4.2.3.4-1 shows this convergence in S_D for succinic acid across this set of solvent shell explicit configurations.

Table 4.2.3.4-1 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for succinic acid (exptl $pK_a = 4.1_6$).

Cluster assignment		pK_a	ΔpK_a
S_D(III+I)			
HA	**A^-**		
$S_C^*[Q_{a1}Q_{br}Q_{a2};Q_{a2}]$	**$S_C^*[Q_{a1}Q_{e1}Q_{e2};Q_{a2}]$**	**5.0_4**	**-0.8_8**
$S_C[Q_{a1}Q_{a2}Q_{e1};Q_{a2}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2};Q_{a2}]$	4.0_0	$+0.1_6$
S_D(III+II)			
HA	**A^-**		
$S_C[Q_{a1}Q_{br}Q_{a2}:Q_{a1}Q_{a2}]$	$S_C[Q_{a1}Q_{e1}Q_{e;}Q_{a1}Q_{a2}]$	4.9_4	-0.7_8
S_D(III+III)			
HA	**A^-**		
$S_C[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{e1}Q_{e2};Q_{a1}Q_{br}Q_{a2}]$	4.9_5	-0.7_9

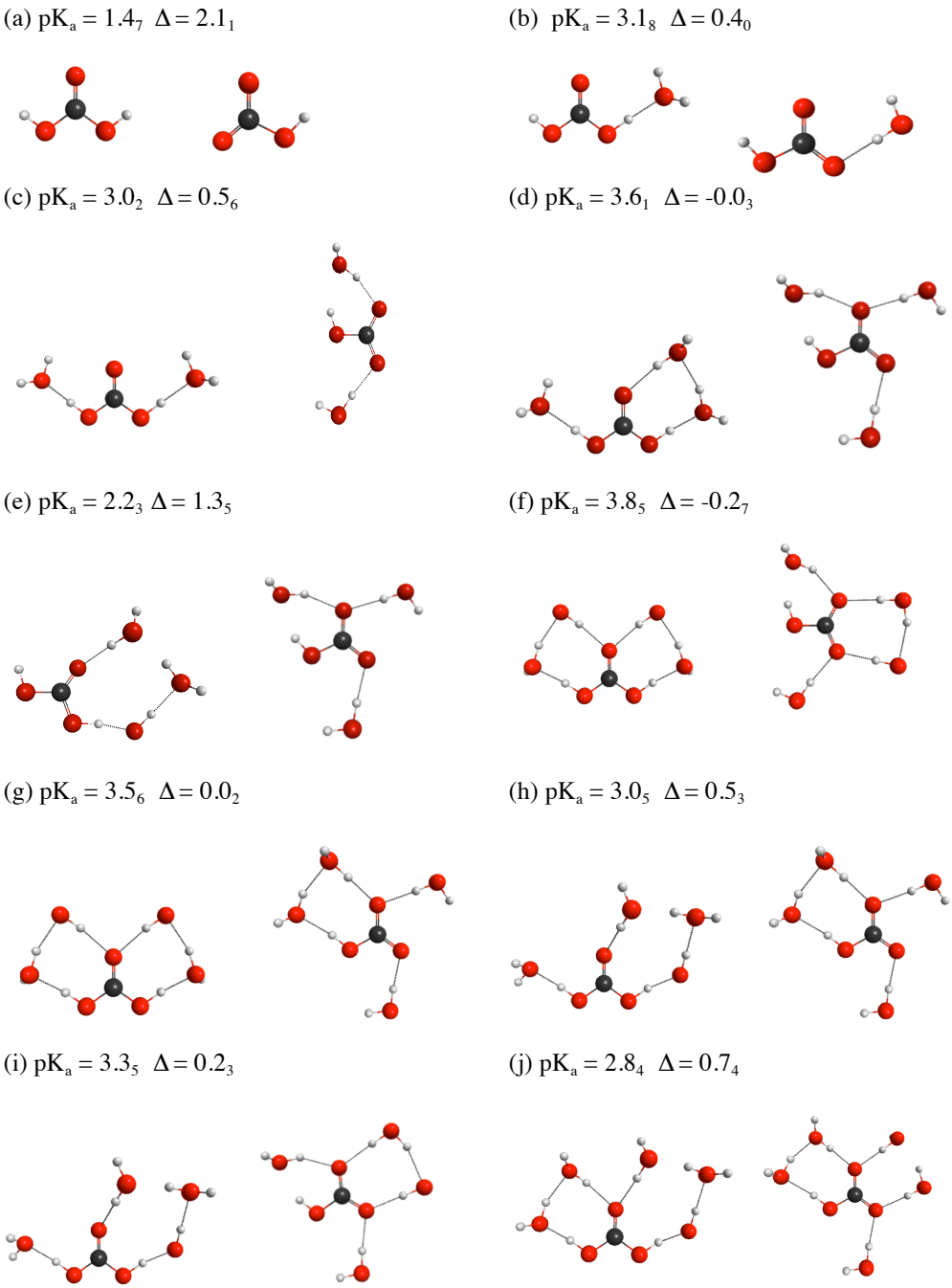


Figure 4.2.2.4.2 B97-D/6-311+G(2d,p) DSES-CC-COSab pK_a as a function of solvation degree (S_D) and solvation sites (Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, Q_{br},Q_{s1}, Q_{s2}, Q_{br-s}) for carbonic acid and associated anion: (a) $S_D(0)$; b) $S_D(I)$, $S_C[Q_{a2}]$:$S_C[Q_{a2}]$; (c) $S_D(I+I)$:$S_D(II)$, $S_C[Q_{a2};Q_{s2}]$:$S_C[Q_{e1}Q_{e2}]$; (d) $S_D(II+I)$:$S_D(III)$, $S_C^*[Q_{a1}Q_{a2};Q_{s2}]$:$S_C[Q_{a1}Q_{e1}Q_{e2}]$; (e) SD(III), $S_C[Q_{a1}Q_{br}Q_{a2}]$:$S_C[Q_{a1}Q_{e1}Q_{e2}]$; (f) $S_D(III+I)$:$S_D(IV)$, $S_C^*[Q_{a1}Q_{a2}Q_{e1};Q_{s2}]$:$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$; (g) $S_D(III+I)$, $S_C^*[Q_{a1}Q_{a2}Q_{e1};Q_{s2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2};Q_{s2}]$; (h) $S_D(III+I)$, $S_C[Q_{a1}Q_{br}Q_{a2};Q_{s2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2};Q_{s2}]$; (i) $S_D(III+I)$:$S_D(IV)$, $S_C[Q_{a1}Q_{br}Q_{a2};Q_{s2}]$:$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$; (j) $S_D(IV+I)$, $S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1};Q_{s2}]$:$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s2}]$.

The set of dicarboxylic acids was extended further to consider more complex bridges than the simple alkyl linkage between the carboxylic acid units. In this category, fumaric and maleic acids, C(O)OH-CHCH-C(O)OH, have intervening unsaturated units in the trans- and the cis- conformations, respectively, terephthalic acid has an intervening aromatic ring, C(O)OH-Ar-C(O)OH, and cyclohexanedicarboxylic acids has an intervening saturated ring unit, C(O)OH-C_6H_{10}-C(O)OH.

Initial predictions for fumaric acid using the preferred S_D(III+1) resulted in an overestimation of the first pK_a value by 1.1_3 units (exptl. $pK_a=3.0_3$). A comprehensive DSES-CC analysis for this acid was therefore conducted (Table 4.2.3.4-2 and Appendix A). Comparing to succinic acid as the unsaturated analogue, the unsaturated bond and second carboxylic group offer substantial stabilization of the charge of the anionic species of fumaric acid substantially lowering the pK_a. Comparison across the series S_D(III+I), S_D(III+II), and S_D(III+III), shows a convergence of results, with S_D(III+III) providing a balanced explicit distribution, and prediction of a pK_a value on the edge of the tolerance limit with respect to the experimental value. The cis isomer, maleic acid has a significantly lower pK_a due to stabilization of the anion through formation of an intramolecular hydrogen bond between the two carboxylic groups in this conformation. The predicted pK_a value using the preferred S_D(III+1) is 2.5_7, which is within an acceptable tolerance of the experimental value.

Terephthalic acid, a para-substituted benzoic acid, is analogous to the para-substituted derivatives in predictive set B. As an electron withdrawing substituent, the COOH-Ar- substituent is expected to make the acid somewhat more acidic acid than benzoic acid. Results using the preferred S_D(III+I) shows a predicted pK_a, value of 4.0_7, which is within the defined tolerance of the experimental value (-0.5_6), and, more acidic than benzoic acid (calcd. 4.7_0). Finally, in the case of cyclohexanedicarboxylic acid, one finds that the preferred S_D(III) with one additional substituent shell explicit solvation (i.e., S_D(III+I), provides a pK_a value within the tolerance limit, at 4.9_6 (exptl: 4.1_8).

The collective expanded predictive set C of pK_{a1} values with 9 diacids using the transferable S_D(III) configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$ with 3 substituent explicit solvents, together with carbonic acid, are reported in Table 4.2.3.2-1

(additional details in Appendix A). The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.71 (std. dev. 0.27).

Table 4.2.3.4-2 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model results for fumaric acid (exptl $pK_{a1} = 3.0_3$).

Cluster Assignment		pK_a	ΔpK_a
$S_D(0)$			
$S_D(0)$		2.7_2	0.3_1
$S_D(I)$			
HA	**A^-**		
$S_C[Q_{a2}]$	$S_C[Q_{a2}]$	3.6_1	-0.5_8
$S_D(I+I)$			
HA	**A^-**		
$S_C[Q_{a2};Q_{a2}]$	$S_C[Q_{a2};Q_{a2}]$	3.9_3	-0.9_0
$S_D(II+II)$			
HA	**A^-**		
$S_C^*[Q_{a1}Q_{a2};Q_{a1}Q_{a2}]$	**$S_C^*[Q_{a1}Q_{a2};Q_{a1}Q_{a2}]$**	**3.8_2**	**-0.7_9**
$S_C[Q_{a1}Q_{a2};Q_{a1}Q_{a2}]$	$S_C[Q_{a1}Q_{a2};Q_{a1}Q_{e2}]$	4.2_0	-1.1_7
$S_D(III+I)$			
HA	**A^-**		
$S_C^*[Q_{a1}Q_{br}Q_{a2};Q_{a2}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2;}Q_{a2}]$	4.1_6	-1.1_3
$S_C[Q_{a1}Q_{a2}Q_{e1};Q_{a2}]$	$S_C^*[Q_{a1}Q_{e1}Q_{e2;}Q_{a2}]$	2.8_2	$+0.2_1$
$S_D(III+II)$			
HA	**A^-**		
$S_C[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{a2}]$	$S_C[Q_{a2}Q_{e1}Q_{e2};Q_{a1}Q_{a2}]$	3.7_1	-0.6_8
$S_D(III+III)$			
HA	**A^-**		
$S_C^*[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	$S_C[Q_{a1}Q_{e1}Q_{e2;}Q_{a1}Q_{br}Q_{a2}]$	3.7_9	-0.7_6
$S_C^*[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	$S_C^*[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	3.7_8	-0.7_5

4.2.3.5 Dicarboxylic Acids, Second Protonation States

The remaining point to be addressed in this section involves the ability of the DSES-CC model to predict multiple acidic protons. In particular, the second acid dissociation constants, pK_a^2, are of interest for the class of dicarboxylic acids, as also recently explored in the literature. [71b] In general, one expects the second protonation state in water to be much weaker (larger pK_a values), since it is more difficult to remove a proton from an anion than from an uncharged molecule. However, the structure of the intervening R group of the COOH – R – COOH will be important in determining the relative acid strength of the remaining proton. In particular, one

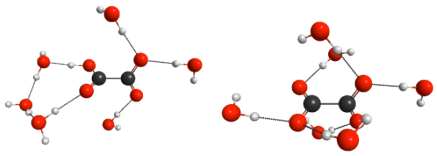
expects that, as the distance between the two carboxylic units increases, the acidity of the second proton to increase.

For all of the diprotic acids except carbonic acid, prediction of pK_a^2 is achieved within the tolerance limit with the preferred DSES-CC configuration around both carboxylic(ate) groups (Figure 4.2.3.5.1). In all cases, pK_a^2 is indeed less acidic than pK_a^1. In particular, oxalic acid has a predicted pK_a^2 that is considerably less acidic than pK_a^1 due to the fact that the second acid proton is held more tightly via an intra-molecular hydrogen bond, as facilitated by the proximity of the carboxyl units.

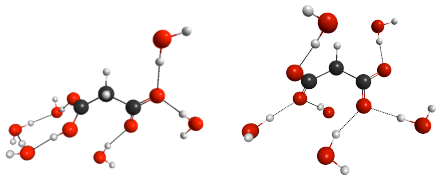
Prediction of pK_a^2 for carbonic acid is considered a special case just as in prediction of pK_a^1, which due to its small size technically has only principal solvent sites (Figure 4.2.3.4.1). As was done for the assignment of S_D for pK_a^1, it is more instructive to refer to the sum of the explicit molecules, rather than the components. The findings from defined sector model study of the training sets reveals the second deprotonation reaction, from carbonate to bicarbonate, to be quite sensitive to explicit placement. However, a converged result is found with a total of five explicit solvent molecules, as shown in Figure 4.2.3.5.2(d).

The collective predictive set C of pK_a^2 values for the 9 diacids using the transferable S_D(III) configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$ together with 3 substituent explicit solvents, and the special case of carbonic acid, are reported in Table 4.2.3.2-1 (additional details in Appendix A). The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.35 (std. dev. 0.29).

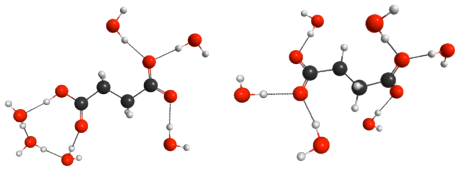
(a) $pK_a = 4.6_2$ $\Delta = -0.4_3$



(b) $pK_a = 5.7_0$ $\Delta = -0.0_1$



(c) $pK_a = 5.5_1$ $\Delta = 0.1_0$



(d) $pK_a = 5.8_5$ $\Delta = -0.4_4$

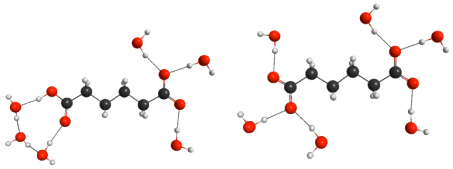
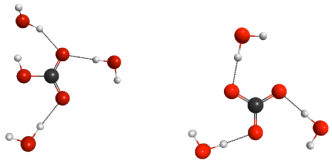
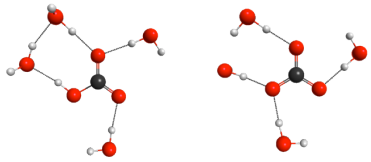


Figure 4.2.3.5.1 B97-D/6-311+G(2d,p) DSES-CC-COSab pK_a for the second deprotonation reaction with the preferred configuration S_D(III+III), $S_C(Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{e1}Q_{e2}]:S_C[Q_{a1}Q_{e1}Q_{e2};Q_{a1}Q_{e1}Q_{e2}]$, for both carboxylic/carboxylate groups of (a) oxalic, (b) malonic, (c) succinic and (d) adipic acids.

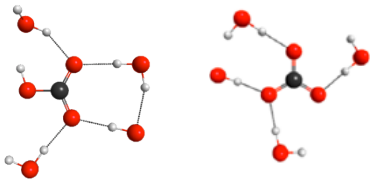
(a) $pK_a = 13.9_3$ $\Delta = -3.3_3$



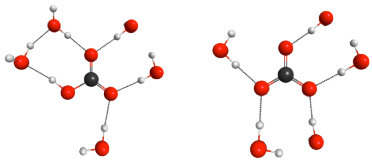
(b) $pK_a = 13.1_9$ $\Delta = -2.5_9$



(c) $pK_a = 12.9_0$ $\Delta = -2.3_0$



(d) $pK_a = 10.9_5$ $\Delta = -0.3_5$



(e) $pK_a = 8.9_8$ $\Delta = 1.6_2$

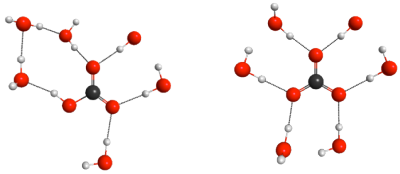


Figure 4.2.3.5.2 B97-D/6-311+G(2d,p) DSES-CC-COSab pKa as a function of solvation degree (S_D) and solvation sites (Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, Q_{br}, Q_{s1}, Q_{s2}, Q_{br-s}) for carbonate and associated anion, bicarbonate: (a) S_D(III):S_D(II+I), $S_C[Q_{a1}Q_{e1}Q_{e2}]:S_C[Q_{a1}Q_{e2};Q_{s2}]$; (b) S_D(III+I):S_D(II+II) $S_C^*[Q_{a1}Q_{e1}Q_{e2};Q_{s2}]:S_C[Q_{a2}Q_{e1};Q_{s1}Q_{s2}]$; (c) S_D(IV):S_D(II+II), $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]:S_C[Q_{a1}Q_{a2}Q_{e2};Q_{s1}]$; (d) S_D(IV+I):S_D(III+II), $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s2}]:S_C[Q_{a1}Q_{a2}Q_{e2};Q_{s1}Q_{s2}]$; (e) SD(IV+II), $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s2}Q_{br-s}]:S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s1}Q_{s2}]$.

4.2.3.6 Alternative S_D(III)'s

Across all systems, a preferred S_D(III) with $S_C[Q_{a1}Q_{br}Q_{a2}]$:$S_C[Q_{a1}Q_{e1}Q_{e2}]$, together with the inclusion of 1-3 explicit solvents in the substituent shell where warranted, appears to satisfy pK_a value prediction within the tolerance limits set out; however, one might expect other possibilities could exist. The key is that any 'preferred' S_C needs to be transferable among a large set of structures and within kT of the thermodynamic minimum, in order to be a faithful representation of the ensemble. For example, one can find a second S_D(III) with the same anion configuration as the preferred anion S_C, but with an alternative acid configuration of $S_C[Q_{a1}Q_{a2}Q_{e2}]$, which also provides excellent prediction of pK_a values. However, while in several cases the new acid configuration is < 0.5 kcal/mol of the preferred acid configuration, there are also several cases where the difference is quite large (e.g., nearly 2 kcal/mol). As such, this alternative S_D(III) does not appear to be a transferable S_D(III) (see., e.g., Appendix A). In the set of acids considered in this study, only one 'preferred' S_D was found, that being $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, and, across the entire set of acids considered, this provided pK_{a1} predictions with MAE of 0.50 (std. dev. 0.28).

4.2.3.7 Higher Degrees of Solvation – Substituent Shell

While generally, within a any particular acid, prediction of pK_a value converged towards the experimental value with principal and secondary explicit solvation sites represented by a 'preferred' S_D(III), one might question whether higher degrees of principal solvation show convergence of predicted pK_a, given that four principal and one secondary explicit solvent sites are present in the carboxyl unit (cf. Figure 4.2.1.1); however, consideration of S_D(IV) was already observed to result in unsatisfactory results for the training set.[35] In this work, a further look into both S_D(IV) and S_D(V) was undertaken for a larger grouping of carboxylic acids (see Supplementary Information provided), to explore more fully whether the S_D(IV) results are anomalous or whether the carboxylic and carboxylate systems are always fully satisfied with S_D(III) in the preferred configuration, $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$.

In all cases considered, prediction of pK_a with a fully saturated solvation shell (i.e. occupation of all 4 principal and 1 secondary solvent sites, Figure 4.2.1.1), $S_D(V)$, as well as $S_D(IV)$ is quite poor and well outside the target tolerance (e.g., Table 4.2.3.7-1 and Appendix A). The question then arises as to why the higher degrees of solvation, $S_D(IV)$ and $S_D(V)$, generally provide poor representations of solution state of carboxylic acids. One might presume that, when the addition of explicit solvent molecules disturbs the "natural" charge distribution of the solute, the predicted pK_a value will be out of the acceptable range of accuracy. It appears that $S_D(IV)$ and higher degrees of solvation tends to overcrowd the solute systems with more directed interaction in the first solvation shell than would be realistic in a dynamic solution environment. Consequently, the additional explicit solvents begin to constitute the bulk, which not only introduces further challenges but also does not provide accurate pK_a prediction. On the other hand, it is conceivable, that these results indicate a fundamental inadequacy in the continuum model approach itself, which is a subject of our future investigations.

In the context of the present DSES-CC model, one can assert that accurate predictions of pK_a for a general carboxylic acid can be realized using the identified 'preferred' explicit solvent network, $S_D(III)$ with $S_c[Q_{a1}Q_{br}Q_{a2}]$: $S_c[Q_{a1}Q_{e1}Q_{e2}]$. This degree of solvation appears to adequately capture the principal, secondary, and substituent shell directed interactions between solute and solvent, with the continuum model capturing the essentials of the bulk.

Table 4.2.3.7-1 B97D/6-311+G(2d,p) Direct-sector explicit solvent in continuum model $S_D(V)$ results for acetic acid (exptl $pK_a = 4.7_6$) and formic acid (exptl $pK_a = 3.7_7$).

$S_D(V)$ Cluster		pK_a	ΔpK_a
Acetic Acid			
HA	**A^-**		
$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	2.5_1	2.2_5
Formic Acid			
HA	**A^-**		
$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	0.7_4	3.0_3

4.2.4 Conclusions

One of the most fundamental reactions in chemistry and biochemistry involves the protolytic reaction of acids and bases, as illustrated by the volume of natural and synthetic organic compounds with acidic or basic functionality. Determining systematic effects of polar substituents on ionization of acids enables establishment of fundamental structure/reactivity relationships. Theoretical prediction of pK_a has been quite challenging and tends to vary widely in chemical accuracy depending on methodology and class of compounds. In particular, for continuum models, a significant challenge has been inclusion of explicit first shell solvation interactions, necessary for accurate prediction of pK_a. The DSES-CC model has been presented as an important step for determining explicit solvation in the first solvation shell. The model has been demonstrated for prediction of both pK_a^1 and pK_a^2 values across a broad range of carboxylic acids, a relatively challenging class of functionality.

In the relative comparison of acid strengths among a series of carboxylic acids, entropy factors are not considered, but are found to make only minor contribution. The relative translational and rotational degrees of freedom between acid and anion are similar for all acids being compared, so that enthalpy factors become the most important factor for prediction of the relative acidities.[61a] In this way, a straightforward approach using only the continuum model plus the appropriate defined-sector model is found to be needed for prediction of acid dissociation constants. Through careful consideration of solute solvent surfaces, the model has enabled generalizations that indicate number and conformation of explicit solvent molecule networks for classes of solutes and associated functionality. For the class of carboxylic acid structure (32 acids, including 9 dicarboxylic acids), a 'preferred' network conformation, consisting of S_D(III), with $S_C^*[Q_{a1}Q_{br}Q_{a2}]$:$S_C^*[Q_{a1}Q_{e1}Q_{e2}]$ plus substituent explicit solvation when necessary, is found to provide pK_a within the tolerance set out, with a MAE of 0.50 pK units (0.7 kcal/mol) accuracy. Moreover, the model shows equal reliability for prediction of pK_{a2} values of dicarboxylic acids.

Future studies will investigation a) the general applicability of the DSES-CC model for other classes of functionality, b) ways to automate the method for S_D/S_C choice and solvent placement, and c) the fundamental nature of the transition from higher

degrees of explicit solvation to the continuum model. Extension of the DSES-CC model for other functionality formally requires designation of principal and secondary explicit solvation sites around the relevant functional groups (e.g., amine, alcohol, carbon acid, etc), as in Figure 4.2.1.1 for carboxylic acid functionality. The present study demonstrates how the DSES-CC model addresses other functionality through the treatment of the substituent shell component of the carboxylic acid sector model. In this way, the DSES-CC sectors for a variety of functionality are illustrated. Future studies should detail the different degrees and configurations of solvation, and preferred solvent networks for other classes of functionality.

4.3 Statistical Analysis

Regression analysis, briefly touched upon in Chapter 3.2.3, can be employed both to correct calculated values to obtain higher agreement with experimental values and to assess the success/failure of the particular computational procedure. The linear fit method follows the relationship,

$$pK_a = A\frac{\Delta G_{diss}}{RTln(10)} + B \qquad (4.3\text{-}1)$$

where $\frac{A}{RTln(10)}$ is the slope of a line fit of experimental values of pK_a against the calculated pK_a values, and B represents the –log of the concentration of the solvent, which is in this case is water, and under standard conditions is equal to -1.74.[46] Theory demands a slope of unity, however, to date has not yet been achieved with any computational methods for pK_a prediction. Importantly, improvements to what has been termed "the slope problem" have been observed with the addition of explicit water molecules[45d]. Klamt and coworkers, in a COSMO-RS pK_a study, addressed the issue of slope in detail, considering ΔG_{diss} into four contributions: dielectric energy of the anion, dielectric energy of the neutral compound, gas-phase energy difference, and chemical potential difference arising from the COSMO-RS model.[46] Carrying out a multi-linear regression, however, was unable to isolate any single factor as the main cause of the divergence from expected thermodynamic equivalence.[46] While they found their COSMO-RS method significant in improving the slope from 50% to 58%

of the theoretical value, they concluded that, "the experimental pK_a-scale does not correspond to the free energy of dissociation in infinite dilution of an acid in pure water and probably even so in other solvents."[46] Their conclusion draws into question the entire computational strategy of calculating the pK_a from the free energy of dissociation of one molecule of the acid, which is based on the assumption of a dilute solution. Whilst this conclusion cannot be ruled out, there are still a number of deficiencies that need to be addressed before one would reach that line of reasoning. In fact in a later study, Klamt combines COSMO-RS with a cluster continuum approach and finds again and improvement in the slope.[45a] This is consistent with the other statistical analyses reported on the use of continuum cluster / implicit-explicit methodologies.[45d, 48b, 49]

In order to further investigate these ideas, an analysis was carried out across all systems considered in the former pK_a studies, to assess the accuracy of our underlying model. Table 4.3-1 shows a number of groupings, along with a final analysis of all the systems considered in both of our published pK_a studies. The groupings follow the categorization outlined in 'Defined-Sector Explicit Solvent in Continuum Cluster Model for computational prediction of pK_a: Consideration of secondary functionality and higher degree of solvation', namely, Predictive set A, predictive set B, predictive set C pK_a^1 and pK_a^2, the entire set of pK_a^1, and all systems.

Unlike the work of Klamt using the COSMO-RS method[46], the data presented here considers solvation energies corrected by zero point energy, $E_{S,ZPE}$, and not full free energies of solution as $\Delta G = \Delta H - T\Delta S$. Calculation of total free energies, ΔG, including entropic contributions in particular, will be returned to in the next chapter. Determination of pK_a from calculated $\Delta E_{S,ZPE}$[3] values and running a correlation against experimental pK_a values produces results as show in Figure 4.3.1. The slope of the pK_a^1 values of all the systems is found to be 0.75 (Figure 4.3.1(a)). This is a vast improvement over the slope of Klamt of 0.58.[46] Whilst higher slopes have been achieved (Adam reports a slope of 0.93[48b] and Kelly et al. of 0.86[45d]), it is important to compare to the completely implicit methodology with the same computational strategy. With the methodology described in this work, without explicit water

[3] $\Delta E_{S,ZPE} = E(A^-)_{S,ZPE} + E(H^+)_{s,exptl.} - E(AH)_{S,ZPE}$

molecules, a slope of 0.31 was found (Figure 4.3.2). This demonstrates the significance of the DSES-CC method for placement of water molecules to study the pK_a of carboxylic acids. The R^2 values for the implicit implementation alone and DSES-CC method (pK_a^1 only) are both 0.85, showing that a line fit does offer an appropriate method for prediction in this case. With the known deficiencies in regard to the thermochemical aspects of calculating pK_a, one would be skeptical of a slope of 100% with the DSES-CC model alone.

Higher slopes are observed for the set of dicarboxylic acids with the DSES-CC method; with a slope of 0.86 for the pK_a^1 values and a slope of 1 for the pK_a^2 values. The R^2 values for these datasets are 0.97 in both cases. The strong correlations and higher slopes are potentially due to the symmetry of the systems constituting these sets, and therefore the potential for error cancellation.

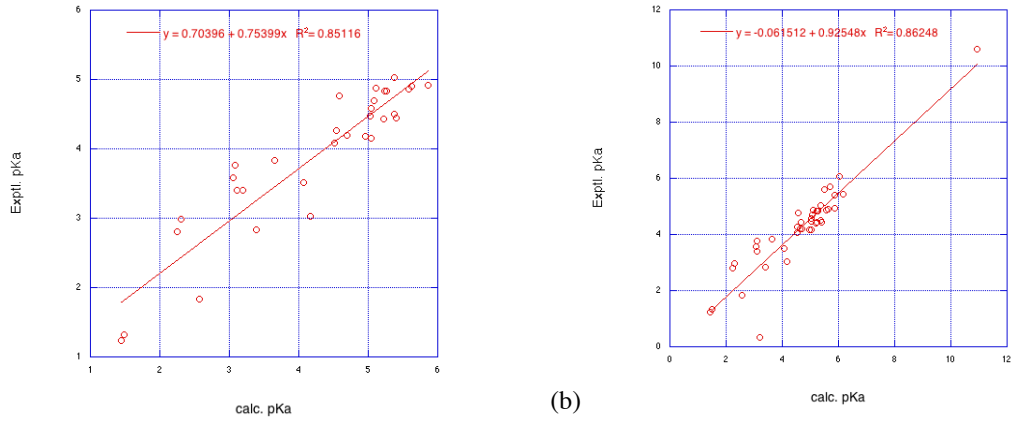


Figure 4.3.1 (a) Correlation of DSES-CC pK_a^1 results against experiment; (b) Correlation of all DSES-CC pK_a results against experiment

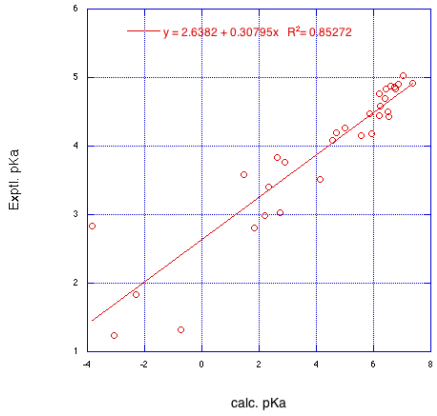


Figure 4.3.2 Correlation of calculated pK_a^1 with pure implicit model compared with experiment

Table 4.3-1 Statistical results for DSES-CC pK_a calculated values against experimental values

	Mean |Δ(exptl-calc)|	SD |Δ(exptl-calc)	Max |Δ(exptl-calc)|	regression analysis: slope	regression analysis: R^2	regression analysis: y intercept
Predictive Set A	0.44	0.23	0.96	0.72	0.89	1.06
Predictive Set B	0.58	0.24	0.94	0.55	0.98	1.69
Predictive Set C: pk_a^1	0.71	0.27	1.13	0.86	0.97	-0.17
Predictive Set C: pk_a^2	0.35	0.29	0.79	1.00	0.97	-0.30
All pk_a^1	0.54	0.26	1.13	0.75	0.85	0.70
All systems*	0.50	0.28	1.13	0.81	0.86	-0.52

* Excluding carbonate because it is the only acid in the set with a pK_a value out past 6 pK units.

4.4 Limitations of the DSES-CC model

The mixed discrete - implicit solvent approach presented in this work, the DSES-CC model, has demonstrated significant improvement over standard CSMs in achieving accurate property predictions, specifically pK_a. Although the model presented here has been demonstrated only for carboxylic acids, the model is extendable to other functionality through modification of the defined-sectors for explicit solvation in the new functionality. In fact, the published, "Defined-Sector Explicit Solvent in Continuum Cluster Model for computational prediction of pK_a: Consideration of secondary functionality and higher degree of solvation", already explores additional functional groups within the substituent shell of the carboxylic acids, that demonstrates such an extension.

The success of the DSES-CC model is presumably due largely to its ability to capture the energetics associated with the important direct interactions between the solute and solvent, which are missing from purely implicit CSMs. However, by doing so it also inherently adds a buffer between the solute and the dielectric continuum or in other words, the explicit representation of the first solvation shell graduates the dielectric value from the solute to that of the bulk solvent. This is likely to be an additional advantage of the continuum cluster methodology.

In the current DSES-CC implementation, and in all continuum-cluster studies reported in the literature, the solvent molecules are treated within the cavity of the solute, instead of in their own solvent shell. The model is therefore treating the 'solute' as an

extended system, and likely therein introduces some error in properties associated with the solute alone. In Figure 4.4.1 the DSES-CC general scheme is overlaid over a Molecular Electrostatic Potential representation of the S_D(III) $S_C[Q_{a1}Q_{e1}Q_{e2}]$ cluster for acetate, both of which are surrounded by a graduated blue solvent sphere to provide a graphical representation of an explicit/implicit solvation model including such a graduated representation of the solvent from the solute through the first/second solvation shell, and out to the bulk continuum.

In the present work, a variety of strategies were explored to address these issues relating to the mega-cluster approach. One would like to maintain the benefit of the explicit water molecules, however not have them present as a super-solute. Two possible strategies in particular were considered towards this goal. The first strategy directly considers essentially a first solvation shell, outside the cavity of the solute alone, where the explicit solvent molecules are placed. This involved enabling new functionality in the programmed COSab model, which essentially separates the treatment of the solute from that of these explicit solvation models. To this end, one can now read in the final N atoms of the geometry input as the explicit water molecules. These explicit water molecules are then excluded during the cavity construction, but their electronic structure is carried out and included in the one-electron Hamiltonian routine.

A second strategy was next developed, which simplifies the first strategy in that it avoids having any nuclear centers outside of the original solute cavity. This second strategy does not include the electronic structure associated with the atoms of the explicit solvent molecules into Hamiltonian, but instead adds a representation of the explicit solvent models by way of associated charge presentation of their cavity surface, which represents their self-consistent interaction with the bulk solvent. This approach also required addition of new functionality to the base model to incorporate the explicit solvent representation in this way. This strategy requires a prior COSab calculation on the explicit water molecules alone (or other solvent molecules in the case that water is not the solvent of choice), to obtain the cavity charges of a single explicit water molecules at a fixed geometry. This representation then had to be incorporated into the self-consistent field iterations through the Hamiltonian representation.

While both of these strategies provide reasonable solutions to the issues of treating the explicit solvent molecules in the first solvation shell, ideally one would like to account for the first solvation shell effects in a fully implicit way, rather than the in this hybrid implicit/explicit manner. Moreover, there are several other issues that still would not be solved given the proposed strategies. For instance, issues regarding treatment of outlying charge correction in these schemes, the incorporation of a graduated dielectric starting in the first solvation shell, and issues of non-electrostatics. As previously outlined, there are currently two strategies for treating outlying charge in COSab in GAMESS, the double cavity OCE correction method and the distributed multipole method. The distributed multipole OCE scheme provides greater flexibility in algorithmic developments however, in these methodology explorations the double cavity method was employed until a bug in the distributed multipole scheme was resolved. The fix for this bug is reported in Appendix B. Due to a number of complexities, a stepwise approach is desirable and Chapter 6 focuses on various methodology developments required in a consideration of the scheme, as depicted in Figure 4.4.1.

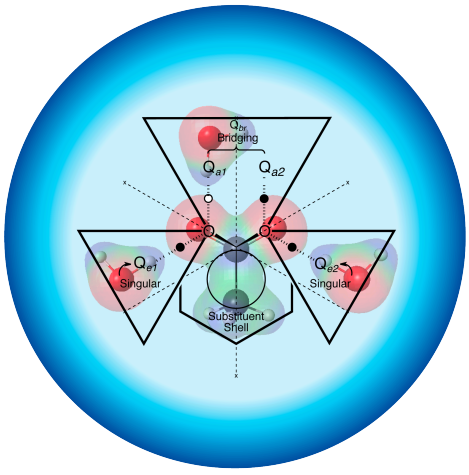


Figure 4.4.1 Illustration of the DSES-CC model, the acetate cluster $S_C[Q_{a1}Q_{e1}Q_{e2}]$ and a rough depiction of a graduated continuum extending from the solute cavity out to the bulk dielectric

5 Second derivative calculations in solvent

5.1 Introduction

Current methodologies in state-of-the-art quantum chemistry calculations are at a level of enabling high precision predictions of energetics and geometries, provided one chooses appropriate levels of theory. As the uncertainties in prediction of geometry and energetics become less and less, other sources of uncertainty become increasingly noticeable. This is particularly true for property predictions. In the case of solvent properties, the previous chapters show that for accurate prediction of aqueous pK_a, it becomes essential to include direct interactions of solvent in the first solvation shell. The DSES-CC model in this way has further improved the predictability of electronic theory for prediction of pK_a, and therefore allows other sources of uncertainty to be addressed.

The second derivative, or Hessian, analysis, can also become a source of error, particularly for properties that are very sensitive to small changes in energetics. In computing solvation properties one also needs to include the contribution of the solvent in the calculation of the second derivatives, creating an additional challenge. In the context of the work in this thesis, the vibrations in the solvated super-cluster also pose an issue, as location of global minima for weakly bound systems is still an unresolved issue,[45a, 61a] primarily due to the soft mode vibrational structure associated with the loose association of solvent to solute.

The important properties obtained from the Hessian calculation for pK_a prediction include the zero-point correction and the statistical thermodynamic parameters associated with the calculation of free energy, as opposed to simply electronic energy or electronic energy plus zero point vibrational energy corrections. In particular, there exists significant controversy concerning the estimations in determination of enthalpy, entropy, and overall free energy in solution. Associated with this is the underlying statistical mechanics associated with translational and rotational motions of solute molecules in a solution environment. Even in experiment, there are issues associated with determination of translational and rotational components to the free energy (e.g.,

enthalpy and entropy) due to the small vibrational motions. The low-frequency vibrational motions of the coupled solute-solvent as well as solvent-solvent components can actually make significant contributions in terms of the entropic contribution to the total free energy,[70] and therefore is an important consideration in the determined pK_a. This chapter firstly explains the standard methodology of a Hessian analysis, so to provide a basis for understanding these issues. Then, the zero-point energy correction is addressed and finally the issues pertaining to the statistical thermodynamic properties are discussed.

5.2 Standard approaches to Hessian analysis in Quantum Chemical Calculations

The frequency, or Hessian, analysis in standard quantum chemistry software involves the calculation of the second derivative of the energy with respect to geometry. This analysis is typically carried out under the harmonic approximation, which assumes a quadratic behavior in the vicinity of the minimum, although anharmonic corrections can be assumed in the analysis. Moreover, the analysis can computationally be carried out by way of analytical methodology, numerical methodology, or semi-numerical methodology. In terms of solvation capabilities within the GAMESS code, the Hessian analysis can only be carried out numerically. The numerical solution follows a finite displacement approach, via calculating the first derivatives for a given geometry, then perturbing the coordinates by a small amount, carrying out a new self-consistent field and gradient analysis at this new geometry, and finally taking the difference between the two gradients divided by the step size. The result of this procedure for all 3N coordinates is the total Cartesian force constant matrix, where the elements are defined as,

$$H_{i,j} = \frac{\delta^2 E}{\delta x_i \delta x_j} \qquad (5.2\text{-}1)$$

Diagonalization of this matrix yields the normal coordinates which are the eigenvalue/eigenvector representations associated with the 3N degrees of freedom of the molecule; the 3 degrees of freedom associated with translation, the 3(2) degrees of freedom associated with rotation, and the 3N-6(5) vibrational degrees of freedom. The

eigenvalues of the matrix are the energies associated with these degrees of freedom and the eigenvectors provide the corresponding motions (vectors) of the 3N coordinates. The 3N harmonic oscillator Schrodinger equations are defined as,

$$\left[-\frac{1}{2}\frac{\partial^2}{\partial Q_i^2}+\frac{1}{2}k_i Q_i^2\right]\psi_i(Q_i)=\epsilon_i\psi_i(Q_i) \qquad (5.2\text{-}2)$$

where k_i are the eigenvalues of the Hessian and Q_i are the displacements along the normal coordinates i, and $\Psi(Q)$ is the nuclear wavefunction,

$$\psi(Q)=\psi_1(Q_1)\psi_2(Q_2)\ldots.\psi_{3N}(Q_{3N}) \qquad (5.2\text{-}3)$$

A number of thermodynamic properties (e.g. entropy and free energy) can be obtained by applying employing statistical thermodynamics and basic quantum mechanical models for the various degrees of freedom, to the normal modes of the Hessian. The particle in the box (translation), rigid rotor approximations (rotation), and harmonic oscillator (vibration) approximations are employed to obtain the partition functions relating to translation, rotation, and vibrations, respectively. As such, one can see that such approximations can also introduce error in the predictions of vibrational modes and associated frequencies, which are used in the determination of zero point energy, as well as the various statistical thermodynamic estimates of enthalpy and entropy. These values can of course be improved on by attention to the level of theory used in the evaluation of any particular property, either in gas phase or solution phase. Very accurate thermochemical calculations using quantum chemical methods would indeed require highly accurately determined electronic energies, anharmonic zero point vibrational energies for enthalpies at 0 K, thermal corrections for enthalpies at different temperatures (e.g., 298.15 K), and corrections due to entropy.

5.3 Zero-point energy corrections

Zero point energy is the motion that a system has at the quantum mechanically determined minimum energy ground state. This is due to the fact that all quantum mechanical systems have a certain innate energy associated with small fluctuations even at 0 K. The zero point energy is calculated as a part of the Hessian analysis as,

$$\varepsilon_{vib}^{H} \cong \frac{\hbar c}{2} \sum_{m} \omega_m \qquad (5.3\text{-}1)$$

where ε_{vib}^{H} in this case refers to the harmonic zero point vibrational energy, ZPVE, ω_m are the computed harmonic vibrational frequencies of the vibrational mode m, in wave numbers, and $\hbar c$ is Planck's constant multiplied by the speed of light, c.

It has been well established that the harmonic approximation typically overestimates the vibrational frequencies by up to 10% depending on the level of theory chosen for the electronic structure theory. With the advent of accurate theories in density functional theory and higher order methods, this typically has been decreased significantly to more like 1-3%. However, even an error of 1% can be important to achieving chemical accuracy,[76] particularly for the prediction of pK_a. One possible improvement is to consider anharmonic corrections.

A rigorous treatment of anharmonicity is quite complex, involving the calculation of higher order (e.g cubic and quartic) force constants and the calculation of the multidimensional potential energy surfaces, where the degrees of freedom increase with molecular size by 3N-6, where N is the number of atoms.[76-77] Empirical corrections have therefore become the most common strategy.[76-77] Scaling factors have shown to be a simple and effective approach to treating the anharmonic effects, and a number of literature studies have determined the appropriate scaling factors for a range of model chemistries.[76-78]

Once such empirical correction has been presented in the literature by the group of Truhlar.[71b] In the case of low frequency vibrational modes, in particular, frequencies below 100 cm^{-1}, the harmonic approximation has been shown to be particularly inadequate. Such modes correlate with weak interactions, as one finds in explicit/implicit solvent systems involving hydrogen bonded clusters. Truhlar and co-workers have proposed a simple empirical correction that involves a) raising all frequencies below 100 cm^{-1} to 100 cm^{-1},[71b, 79] and then scaling all frequencies by a factor that has been determined to be appropriate for the methodology choice. Whilst such a correction has been demonstrated in a few studies,[71b, 79-80] there has been no clear argument presented for fixing of all frequencies below 100 cm^{-1} to 100 cm^{-1}.

In the present work, the empirical correction from the group of Truhlar[71b] was considered to investigate the effect of the proposed correction on the systems considered in our investigation of pK_a in, "Defined-Sector Explicit Solvent in Continuum Cluster Model for computational prediction of pKa: Consideration of secondary functionality and higher degree of solvation." A scaling factor of 0.9904 appropriate for the B97-D functional together with a polarization consistent basis set including d-polarization,[77] was used. The corrected pK_a values vary only slightly from those calculated using harmonic zero point energies, and, importantly, do not change the overall conclusions of the DSES-CC model predictions. Several of the exceptional cases already discussed in the DSES-CC model study, namely cinnamic acid and p-aminobenzoic acid, are shifted further outside the range (e.g. the result for cinnamic acid is 1.26 pK units from experiment, compared with 0.94 without the correction) of the tolerance limit (i.e., 0.74 pK units or 1 kcal/mol), albeit not unreasonably. However, most of the exceptional systems found have challenges unrelated to small vibrational corrections, making it unlikely that the approximate empirical correction as presented is a reliable determinant for driving these values further outside the tolerance limit. More importantly, there are other concerns related to vibrational corrections that would need to be investigated that further influence prediction. As, such, we are at a level of accuracy where empirical corrections can introduce more uncertainty in prediction, and therefore, it is more reliable to put more effort into a proper treatment of anharmonicity including all of the important effects.

Table 5.3-1 Anharmonic correction to ZPVE energies in the calculation of pK_a

Acid	S_D	Exptl pk_a	DSES-CC pK_a	ΔpK_a	DSES-CC_{anharm} pK_a	ΔpK_a
Initial Predictive Set – from paper 1						
acetic	S_D(III)	4.76	4.58	0.18	4.91	-0.15
formic	S_D(III)	3.77	3.09	0.68	3.25	0.52
propanoic	S_D(III)	4.86	5.58	-0.72	5.82	-0.96
isobutyric	S_D(III)	4.88	5.11	-0.23	5.31	-0.43
trimethylacetic	S_D(III)	5.03	5.37	-0.34	5.44	-0.41
chloroacetic	S_D(III)	2.81	2.25	0.56	2.52	0.29
glycolic	S_D(III)	3.84	3.65	0.19	3.80	0.04
benzoic	S_D(III)	4.2	4.7	-0.5	4.79	-0.59
Expanded Predictive Set A – increasing bulk, EWG, unsaturated						
butanoic	S_D(III)	4.83	5.24	-0.41	5.38	-0.55
pentanoic	S_D(III)	4.84	5.26	-0.42	5.42	-0.58
cyclohexanecarboxylic	S_D(III)	4.9	5.62	-0.72	5.99	-1.09
nitroacetic	S_D(III)	1.32	1.49	-0.17	1.72	-0.40
mandelic	S_D(III)	3.41	3.11	0.3	3.19	0.22
acrylic	S_D(III)	4.26	4.55	-0.29	4.88	-0.62
crotonic	S_D(III)	4.69	5.08	-0.39	5.34	-0.65
trans-cinnamic	S_D(III)	4.44	5.4	-0.96	5.70	-1.26
Predictive Set B – aromatic acids						
o-hydroxybenzoic	S_D(III)	2.98	2.3	0.68	2.40	0.58
m-hydroxybenzoic	S_D(III)	4.08	4.52	-0.44	4.77	-0.69
p-hydroxybenzoic	S_D(III)	4.58	5.04	-0.46	5.23	-0.65
p-methoxybenzoic	S_D(III+I)	4.5	5.24	-0.74	5.39	-0.89
p-butylbenzoic	S_D(III)	4.47	5.02	-0.55	5.25	-0.78
p-aminobenzoic	S_D(III+I)	4.92	5.86	-0.94	6.10	-1.18
p-nitrobenzoic	S_D(III)	3.4	3.19	0.21	3.36	0.04
***Predictive Set C – pK_{a1} of* Diacids**						
carbonic	S_D(III+I)	3.58	3.05	0.53	3.15	0.43
oxalic	S_D(III+I)	1.23	1.44	-0.21	0.97	0.26
malonic	S_D(III+I)	2.83	3.39	-0.56	3.62	-0.79
succinic	S_D(III+I)	4.16	5.04	-0.88	4.14	0.02
adipic	S_D(III+I)	4.43	5.23	-0.8	5.39	-0.96
fumaric	S_D(III+III)	3.03	3.78	-0.75	3.79	-0.76
Maleic	S_D(III+1)	1.83	2.57	-0.74	2.81	-0.98
terephthalic	S_D(III+1)	3.51	4.07	-0.56	4.21	-0.70
cyclohexanedicarboxylic	SD(III+1)	4.18	4.96	-0.78	5.17	-0.99
Predictive Set C – pK_{a2} of Diacids						
carbonic	S_D(V)'	10.6	10.95	-0.35	11.05	-0.45

oxalic	S_D(III+III)	4.19	4.62	-0.43	4.60	-0.41
malonic	S_D(II+II)	5.69	5.48	0.21	5.81	-0.12
adipic	S_D(III+III)	5.41	5.85	-0.44	6.07	-0.66
succinic	S_D(III+III)	5.61	5.51	0.1	5.83	-0.22
fumaric	S_D(III+III)	4.44	4.66	-0.22	4.93	-0.49
Maleic	S_D(III+III)	6.07	6.04	0.03	6.26	-0.19
terephthalic	S_D(III+III)	4.4	5.19	-0.79	5.12	-0.72
cyclohexanedicarboxylic	S_D(III+III)	5.42	6.17	-0.75	6.21	-0.79

5.4 Statistical Thermodynamics

At this point, the proposed DSES-CC model has considered only solvation energies corrected by zero point energy. For determination of solution-phase free energies, one needs to consider effects of statistical thermodynamics corrections. However, proper and accurate treatment of such effects for determination of solution related properties are a matter of some controversy in the literature.

One common strategy presented in the literature for determination of free energy of solution is to employ a thermodynamic cycle, as discussed previously in the context for calculation of pK_a. This results in the following equation,

$$G_{soln} = G_{gas} + \Delta G_{solv} + RTln(\frac{RT}{P}) \qquad (5.4\text{-}1)$$

where the final term is necessary for the conversion from gas-phase standard state (defined by T and P) to the solution phase standard state of 1M.[45b]

However the formalism of ΔG_{solv} is very often ascribed to the difference between the solvent phase energy and gas phase energy, taken as,

$$\Delta G_{solv} = E_{soln} - E_{gas} \qquad (5.4\text{-}2)$$

As such the assumption is made that the statistical thermodynamics of the gas phase and solution phase systems are comparable and therefore cancel.[45b] Continuum solvation models that utilize parameterized corrections for treatment of the non-electrostatic effects, or, employ optimized radii for cavity construction, inherently do incorporate, to some degree, thermal corrections because they are fit to achieve

experimental free energies of solvation.[45b] As such, predictions of solvation free energies using equation 5.4-1 and based on these strategies with parameterized non-electrostatic terms may have the consequences of double counting thermal contributions.

The statistical thermodynamics of an ideal gas, and the harmonic oscillator rigid rotor approximation, provide an established route to access gas phase thermodynamic functions (entropy, enthalpy, free energy). In solvent phase however, as has been demonstrated by experimental spectroscopists, interactions with surrounding solvent molecules fundamentally change molecular motion.[9] Translational motion in solution environment is simply the liberational free energy, as still calculated via the particle-in-a-box theory (taking into account same standard-state concentration).[79] While typically rotational and vibrational motion are coupled, rotational motion is typically treated separately in both gas and solution phase. However, solute rotational motion in solution environment is simply librational motion, due to the coupling with the surrounding solvent molecules. This latter term has generally been viewed as a long-standing issue for realistic inclusion into a solvent model. Parameterization of CSMs may be able to absorb many of these small effects,[79] however, it is important to be aware of the deficiencies of the models when considering calculated solution-phase "free energies," under these assumptions. This remains an area that requires further contributions.

6 COSab Development

6.1 Introduction

In order to avoid abrupt changes in dielectric from solute to solvent, as governed by the cavity, it is interesting to consider a distance-dependent dielectric. Cossi and coworkers in fact proposed such a distance-dependent dielectric function within the PCM solvation model approach, as a 'non-homogeneous dielectric'.[81] In their implementation, polarization charges appear not just on the cavity surface but also radially throughout the bulk of the dielectric.[81] The bulk dielectric is subdivided into finite regions by mapping the tesserae of the cavity surface, and volume charges are placed along radial lines originating at the center of the solute charge and passing through the midpoints of the surface tesserae (Figure 6.1.1).[81] These are then treated in the same way as the cavity surface charges. Importantly, any appropriate function can be used to describe the dielectric at distance, r. In the work of Cossi et. al., the non-homogeneous dielectric follows a Block-Walker function, specified as,

$$\varepsilon(r) = \varepsilon_B \exp\left(-\frac{a log \varepsilon_B}{r}\right) \qquad (6.1\text{-}1)$$

where a is the distance of each surface tesserae from the center of charge and, ε_B, is the value of the bulk dielectric.[81]

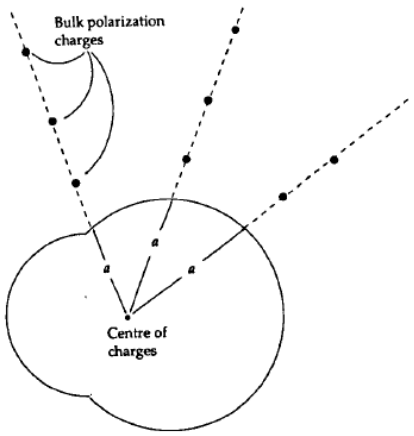


Figure 6.1.1 A depiction of the bulk polarization charges, taken from Cossi et al., 1994[81] with permission from Elsevier.

Examples of distance-dependent dielectric functions are also found in the areas of classical simulations and empirical models (e.g., molecular dynamics, statistical

mehcans, Monte Carlo).[82] In this context, various distance-dependent functions have been employed, such as the Hopfinger model and Langevin functions, which graduate the dielectric value from the cavity values to the bulk solvent.[82a] Generally, distance dependent functions have yielded more accurate descriptions of the electrostatics at short distances from the solute than have constant dielectric values.[82b]

Fattebert and Gygi developed an interesting proposal for distance dependent dielectric capabilities that depends on the electron density of the solute molecule.[83] Their scheme is coupled to an ab initio molecular dynamics method. The distance dependent function they propose involves three parameters; ε_∞, the asymptotic value of the dielectric function (ε_l) as the density tends to zero, ρ_0, the critical density in the middle of the interface solute-solvent, and a tunable variable, β, which changes with the width of the interface.[83] The local model they arrive at is,

$$\frac{\delta \varepsilon_l}{\delta \rho}(r) = \frac{1-\varepsilon_\infty}{\rho_0} \cdot \frac{2\beta \left(\frac{\rho(r)}{\rho_0}\right)^{2\beta-1}}{(1+\left(\frac{\rho(r)}{\rho_0}\right)^{2\beta})^2} \qquad (6.1\text{-}2)$$

Figure 6.1.2 is taken from their publication and shows the shape of the dielectric function as compared to a PCM molecular cavity (dotted circles) for methanol.[83] In a recent effort by Andreussi et al., a revised version of this approach was proposed to solve a number of challenges in the original model of Fattebert and Gygi.[84] They outline the conditions that the dielectric function must obey, specifically,

1) Boundary conditions: the dielectric function should range from a value of 1 (vacuum) inside the molecular cavity to the dielectric value of the bulk solvent.[84]
2) Above certain density thresholds the dielectric should be 1 to avoid "spurious polarization effects due to the interaction of the dielectric with the ion cores".[84]
3) Below certain density thresholds the dielectric should be equal to the bulk dielectric value again to avoid "spurious polarization charges in the bulk of the solvent due to the potential numerical noise in the exponentially vanishing electronic density away from the solute".[84]

4) Smooth function: this is important for their application into plane-wave, periodic electronic-structure code, but may be important in other applications as well.[84]

In their dielectric function, Andreussi et al. introduced a trignometric switching function, the derivative of which is well behaved (equation 6.1-3 & 6.1-4).[84] Importantly the function proposed abolishes the need for any parameterized variables and only requires the dielectric constants and the two density thresholds. [84]

$$\epsilon_{\epsilon_0,\rho_{min}\rho_{max}}(\rho^{elec}) = \begin{cases} 1 & \rho^{elec} > \rho^{max} \\ \exp\left(t(ln\rho^{elec})\right) & \rho^{min} < \rho^{elec} < \rho^{max} \\ \epsilon_0 & \rho^{elec} < \rho^{min} \end{cases} \quad (6.1\text{-}3)$$

$$t(x) = \frac{ln\epsilon_0}{2\pi}\left[2\pi\frac{(ln\rho_{max} - x)}{(ln\rho_{max} - ln\rho_{min})} - \sin\left(2\pi\frac{(ln\rho_{max} - x)}{(ln\rho_{max} - ln\rho_{min})}\right)\right] \quad (6.1\text{-}4)$$

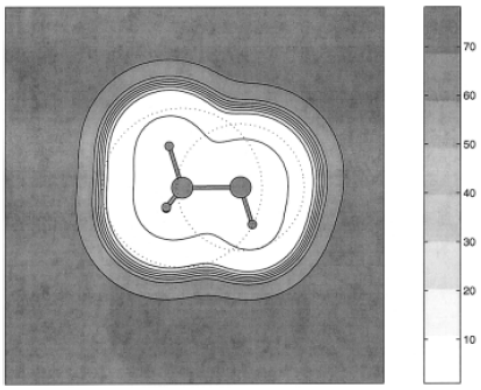


Figure 6.1.2 Contour plot of the dielectric function ε with ϱ_0=0.0004 and β = 1.3 for CH_3OH in a plane containing C-O, taken from Fattebert & Gygi 2001[83] with permission from John Wiley and Sons.

The idea of using a function of the electron density of the solute to attenuate the dielectric is appealing as it offers a physically rigorous means to implement a distance dependent dielectric. In the COSab model in GAMESS, such a distance-dependent dielectric could readily be implemented by replacing the current $f(\varepsilon)$ with $f(\varepsilon) = f(\varepsilon(\rho^{elec}(r)))$, where the dielectric, ε, is a function of the electron density, ρ, at

distance *r* from the atomic center. The functional itself requires further investigation, however the functional used by Andreussi et al. deserves consideration.

There are a number of prerequisite developments that must be investigated before implementation of a distant dependent dielectric. A necessary component is the development of an isodensity cavity construction algorithm that would enable attenuation of dielectric between the various electron density isocontour surfaces. In the following sections a number of the fundamental components of a distance dependent dielectric COSab model are addressed towards this goal.

6.2 Harnessing the Distributed Multiple Algorithm

The matter of outlying charge needs to be considered when proposing a distance dependent dielectric scheme. Currently, the most common procedure (and default in GAMESS) for accounting for the OCE is the double cavity approach. The double cavity OCE approach is a post SCF method and offers a quick route to correcting for the OCE. However, in the drive to create a model with the fewest empirical parameters, several points of the double cavity method should be considered. The first concerns the fact that the distance at which the double cavity is placed is empirically determined to achieve the best correlation with experimental, ΔG_{solv}. As such, the double cavity addresses not only the error related to the amount of solute density beyond the primary cavity, but encompasses other errors associated with the model due to the fit to experiment. Furthermore, the method involves establishing the secondary cavity at a fixed distance from primary cavity for all molecules considered. However, one can imagine particularly diffuse functionality or charge-separated systems, where there still may be tails of the wavefunction penetrating the second cavity. Finally, it is likely that in any implementation of a distance dependent dielectric scheme the treatment of the outlying charge error will need to be reconsidered within the algorithmic scheme.

A second method for treatment of outlying charge area unique to GAMESS is the distributed multipole analysis. This method provides a second representation of the solute potential using a distributed multipole up to hexadecapoles. The difference

between the directly integrated potential determined via the SCF solution of the Hartree Fock equations, and the "corrected" potential determined using the distributed multipole, constitutes the OCE in this analysis. The method is as accurate as the wavefunction method applied in the calculation, and therefore provides a very clean approach for treating the outlying charge error. The only drawback to the method is that the initial implementation is computationally very expensive, prohibiting its use on even medium size molecular systems. Therefore, in order to make further algorithmic considerations using this method, it was desirable to consider strategies to make the algorithm more efficient.

Originally formulated by Stone in 1981, the distributed multipole is calculated by taking the electron density matrix over a basis of Gaussian-type orbitals, with the information for the nuclear charges and positions.[85] When large diffuse basis functions are used the distributed multipoles are calculated by numerical quadrature over a grid of points. The Stone analysis was the first distributed multipole strategy considered for the GAMESS implementation of COSab. However, it was found that results using this model are basis set dependent, and therefore could not provide a reliable approach for prediction of solvation phenomenon. Instead, a second distributed multipole analysis was found, developed by Roger Amos and embedded in the software, CADPAC. For this method, results were found to be much more reliable, but was extremely computationally expensive exactly in the distributed analysis source code in GAMESS (NUMPROP).

The first consideration in addressing the efficiency of the distributed multipole method in GAMESS concerned the fact that the routines involving this analysis were run sequentially. Therefore, the possibility of a parallel implementation of the code was explored, using the available parallel tools already embedded in GAMESS.

The scheme for the distributed multipole outlying charge correction is as follows: The driver routine controlling the algorithmic flow in a cosmo job is COSADD. The keyword option for carrying out a distributed multipole outlying charge method is DMULTI. When the DMULTI option is specified for the outlying charge method, COSADD calls the NUMPROP subroutine, the main driver in the dmulti source file, to

obtain the distributed multipole representation of the electron density for the molecule of interest. The NUMPROP subroutine in turn calls the NUMPRP subroutine, which gets the wavefunction density (DENSGET), and proceeds to carry out the distributed multipole analysis on this density. Initial efforts involved timing various structural components (e.g., DO LOOP structures) in these various subroutines to identify the most time consuming processes in the scheme. It was determined that most of the work for the DMULTI analysis takes place in GRIDMAKE, which has a 5-nested do loop structure that carries out the quadrature for the distributed multipole analysis (Figure 6.2.1).

The DO LOOP structures 20, 30 and 40, were first identified as possible components of large work load as they run over the quadrature parameters NPHI, NTHETA and NR, and are related to the Euler-Maclaurin radial integration, the Gauss-Legendre quadrature, and the phi integration. Further timing tests in these DO structures enabled determination of how large NR, THETA and NPHI loops were getting for varying system sizes. It was ascertained that the loops over NR, NTHETA or NPHI were actually not responsible for consuming lots of time. Rather, timing calls (e.g., TIMIT, in GAMESS), showed that considerable time was amassing around the loop over all atoms, loop 10. Once identified, parallel strategies could be considered using the GAMESS parallel tools already implemented for parallelization, referred to as the Distributed Data Interface (DDI).

To take advantage of the parallelization tools, the COMMON group PAR must be added and appropriate logicals specified as follows,

```
LOGICAL SLB,DLB
LOGICAL GOPARR,DSKWRK,MASWRK
COMMON /PAR   /
ME,MASTER,NPROC,IBTYP,IPTIM,GOPARR,DSKWRK,MASWRK
```

Initialization of the parallel procedure is carried out by setting a counter,
IPCOUNT = ME-1

In the loop structure, the work is partitioned over multiple processors using the following code structure, where X is the reference of the loop,

```
IF (GOPARR) THEN
IPCOUNT = IPCOUNT + 1
IF (MOD(IPCOUNT,NPROC).NE.0) GO TO X
END IF
```

Finally, after the end of the loop structure, the elements from each processor need to be recombined. In this case the elements are all the partial contributions to the molecular multipoles. This is carried out with the GAMESS routine DDI_GSUMF, which requires the arguments, messagetag, arrayname and length as follows,

```
IF (GOPARR) CALL DDI_GSUMF(messagetag,arrayname,length)
```

Two loops over all atoms were targeted, loop 10 as shown in Figure 6.2.1, and a second loop, where a reference had to be added, now loop 4 and again was a loop over all atoms. The final modifications to parallelize the loop over all atoms are shown in Figure 6.2.2.

There are two important metrics to express how much faster a parallel algorithm is as compared with the sequential process; speedup and efficiency. The speedup, S_n, is defined as,

$$S_n = \frac{T_1}{T_n} \qquad (6.2\text{-}1)$$

where, T_1 is the execution time of the sequential algorithm and T_n is the amount of time of the parallel algorithm with n processors.

Efficiency, E_n, is defined as,

$$E_n = \frac{S_n}{n} \qquad (6.2\text{-}2)$$

The efficiency is a measure of how well utilized a set of processors are, and enables determination of the number of processors at which the time required for communication between the processors, exceeds the gained by splitting the load.

```
DO 90 NRTYP1=ILIM1,ILIM2
DO 10 NATOM=1,NATM
DO 20 IR=1,NR
        DO 30 IQ=1,NTHETA
DO 40 IPHI=1,NPHI
```

Figure 6.2.1 5-nested do-loop structure in NUMPRP.

a)

```
      LOGICAL GOPARR,DSKWRK,MASWRK
      LOGICAL SLB,DLB
      COMMON /PAR   / ME,MASTER,NPROC,IBTYP,IPTIM,GOPARR,DSKWRK,MASWRK
```

b)

```
C  initialize parallel
      IPCOUNT = ME-1
c
C
      IF(ISECD1.EQ.7)THEN
      DO 4 I=1,NATM
C  ------GO PARALLEL! --------
         IF (GOPARR) THEN
            IPCOUNT = IPCOUNT + 1
            IF (MOD(IPCOUNT,NPROC).NE.0) GO TO 4
         END IF
C
         DIPX=DIPX+ZAN(I)*C(1,I)
         DIPY=DIPY+ZAN(I)*C(2,I)
         DIPZ=DIPZ+ZAN(I)*C(3,I)
C
```

c)

```
      IPCOUNT = ME-1
C
      NPOINT=0
c
      DO 10 NATOM=1,NATM
         IF (GOPARR) THEN
            IPCOUNT = IPCOUNT + 1
            IF (MOD(IPCOUNT,NPROC).NE.0) GO TO 10
         END IF
C
      NATOMX = NATOM
```

d)
```
C
C*** ATOM CLOSE
   10 CONTINUE
c    ---- sum up all partial contributions of molecular multis
c
      if (goparr) then
      IF (GOPARR) CALL DDI_GSUMF(891,totchg,1)
      IF (GOPARR) CALL DDI_GSUMF(890,ATCHRG,natm)
      IF (GOPARR) CALL DDI_GSUMF(881,dipx,1)
      IF (GOPARR) CALL DDI_GSUMF(882,dipy,1)
      IF (GOPARR) CALL DDI_GSUMF(883,dipz,1)
      IF (GOPARR) CALL DDI_GSUMF(883,atdip,3*natm)
      IF (GOPARR) CALL DDI_GSUMF(884,tquad,9)
      IF (GOPARR) CALL DDI_GSUMF(884,tatquad,3*3*natm)
      IF (GOPARR) CALL DDI_GSUMF(885,soct,7)
      IF (GOPARR) CALL DDI_GSUMF(888,atoct,7*natm)
      IF (GOPARR) CALL DDI_GSUMF(886,shex,9)
      IF (GOPARR) CALL DDI_GSUMF(889,athex,9*natm)
      endif
   90 CONTINUE
```

Figure 6.6.2.2 All code modifications (a – d) necessary for the parallelization of dmulti

The parallel and sequential implementations of the DMULTI option were computed for systems with 5, 8, 12, 18 and 24 atoms. Resulting data is shown in Table 6.2-1. Given that other parts of the GAMESS code are also parallelized, it was important to consider the final speedup specific to the changes described in this chapter as the difference between a fully parallelized code, S_n^{all}, and a code parallelized except for the dmulti algorithm, $S_n^{all\_except\_dmulti}$, as,

$$S_n^{dmulti} = S_n^{all} - S_n^{all\_except\_dmulti} \qquad (6.2\text{-}3)$$

The speedup values without the added parallelization for the DMULTI capability show only minor speedups from a completely sequential process, and as the molecular size increases to 18 atoms, one finds is no advantage over the completely sequential process. Up until systems of 12 atoms there is an advantage in running parallel up until 32 processors, after which efficiencies of below 1 are observed. Importantly, with the parallelization of the DMULTI scheme, significant increases in speedup are observed. Speedups of up to 5.6, 6.9, 11.8, 12.9. and 7.8 are observed for the 5, 8, 12, 18 and 24 atoms systems on 16, 32, 64, 64, and 64 processors respectively.

Table 6.2-1 Parallelization results of dmulti

		Completely sequential process	All parallel except dmulti			All parallel with new dmulti parallel algorithm			
Molecule (and # atoms)	# processors	Total CPU time (s)	Total CPU time (s)	S_n	E_n	Total CPU time (s)	S_n	E_n	S_n^{dmulti}
Methane (5)	4	112.9	107.3	1.05	0.38	73.6	1.53	0.38	0.48
	8		91.4	1.24	0.32	44.3	2.55	0.32	1.31
	16		75.2	1.50	0.35	20.2	5.59	0.35	4.09
	32		82.5	1.37	0.15	23.5	4.80	0.15	3.44
	64		99.5	1.13	0.09	20.7	5.45	0.09	4.32
	128		129.1	0.87	0.04	25.1	4.50	0.04	3.62
Ethane (8)	4	519.2	463.6	1.12	0.45	286.3	1.81	0.45	0.69
	8		402.4	1.29	0.45	145.4	3.57	0.45	2.28
	16		524.5	0.99	0.41	80	6.49	0.41	5.50
	32		427.2	1.22	0.22	75	6.92	0.22	5.71
	64		1035.4	0.50	0.10	80.1	6.48	0.10	5.98
	128		1342	0.39	0.05	90	5.77	0.05	5.38
Benzene (12)	4	3681.9	2997.9	1.23	0.49	1878.9	1.96	0.49	0.73
	8		2665.2	1.38	0.49	946.6	3.89	0.49	2.51
	16		2600.9	1.42	0.38	611.1	6.03	0.38	4.61
	32		4153	0.89	0.24	486.4	7.57	0.24	6.68
	64		7729.7	0.48	0.18	311.7	11.81	0.18	11.34
	128		8701.4	0.42	0.05	531.8	6.92	0.05	6.50
Naphth-alene (18)	4	14258	17851.2	0.80	0.40	8978.6	1.59	0.40	0.79
	8		18769	0.76	0.49	3632.7	3.92	0.49	3.17
	16		14528.1	0.98	0.36	2475	5.76	0.36	4.78
	32		14761.9	0.97	0.24	1852.3	7.70	0.24	6.73
	64		28380.1	0.50	0.20	1103.8	12.92	0.20	12.41
	128		34534.4	0.41	0.07	1554	9.18	0.07	8.76
Anthr-acene (24)	4	35741.1	31363.1	1.14	0.11	79812	0.45	0.11	-0.69
	8		28221.6	1.27	0.38	11867.3	3.01	0.38	1.75
	16		41920.3	0.85	0.31	7191.9	4.97	0.31	4.12
	32		49975	0.72	0.20	5484.9	6.52	0.20	5.80
	64		81210.7	0.44	0.12	4570.8	7.82	0.12	7.38

6.3 No outlying charge correction

In the current implementation of COSab in GAMESS, the user either specifies an outlying charge correction method when running a calculation, or, the program defaults to the double cavity method for determination of outlying charge. For development purposes, it is necessary to include an option for disabling any outlying charge correction. To enable this functionality a third OUTCHG option was created, OCENON, an included in the appropriate DATA block, as,

```
DATA DMULTI,DBLCAV,OCENON/8HDMULTI  ,8HDBLCAV  ,8HOCENON  /
```

This new option must also be initiated, which was carried out by including the following code section,

```
      OK = .FALSE.
      IF(OUTCHG.EQ.BLANK) OUTCHG = DBLCAV
      IF(OUTCHG.EQ.OCENON)   OK=.TRUE.
      IF(OUTCHG.EQ.DMULTI)   OK=.TRUE.
      IF(OUTCHG.EQ.DBLCAV)   OK=.TRUE.
```

For the most part, the existing IF blocks relating to the DBLCAV and DMULTI procedures are specific to these two outlying charge correction schemes. However, there are two locations in the subroutine DECORR where the conditional of DBLCAV required extending to include the new option OCENON, since these routines are required in all cases except in the case of the DMULTI option. These are:

```
      IF (OUTCHG.EQ.DBLCAV.OR.OUTCHG.EQ.OCENON) THEN
       CALL COSCHOL2(ABCMAT,QSCNET,QVPOT,NPS,1.0D00)
      ENDIF
```

And,

```
        IF (OUTCHG.EQ.DBLCAV.OR.OUTCHG.EQ.OCENON) THEN
        DO 173 I=1,NPS
          SE = SE + QVPOT(I)*QSCNET(I)
 173 CONTINUE
        ENDIF
```

In the subroutine COSOCE, code was added at the top of the subroutine to skip the subroutine if the OCENON was chosen. This involved adding a new reference at the end of the subroutine and then calling that reference if OUTCHG=OCENON was selected.

6.4 Cavity Surfaces

In moving towards the consideration of a distance dependent dielectric, it was of interest reexamine the general options for construction of a cavity surface in GAMESS. The current cavity construction strategy belongs to the category of atomic radii-dependent molecular shaped cavity. Within this category, a more systematic view was taken for choice of atomic radii, including new options for choice of radii, and corrected view of existing radii choices. These are discussed in the following section, Chapter 6.4.1. An implementation for an isodensity cavity surface, where the cavity is constructed at a uniform electron density value, is also investigated, as discussed in section 6.4.2.

6.4.1 Van der Waal radii cavities

Molecular shaped cavities generated from an overlapping spheres method either rely on parameterized radii or van der Waal radii. Small differences in the chosen radii can result in significant changes in resulting solvation energies calculated. To demonstrate this point, consider the Born algorithm. In this simple spherical shaped cavity model, a radius for a monatomic ion of 2.5 Å instead of 2 Å would result in a 20% difference in the solvation energy.[86]

While there are some arguments that the vdW cavity is a somewhat arbitrarily decided point at which to define the cavity boundary, it maintains chemical validity theoretically, as it should represent the region where solute atoms can interact with solvent atoms. However, the entire concept of the van der Waals radii assumes that there is a defined radius specific to each atom, regardless of the surrounding

environment.[87] Van der Waal radii, designated as such by Pauling because they represent van der Waals interactions between atoms,[88] constitute a number of different data sets, calculated from varying methodologies. Pauling's radii are calculated from taking the average contact distances between non-bonded atoms from crystallographic data. Bondi further refined the Pauling's radii, by comparing to evaluations based on thermodynamic and physical properties and also to a set of radii constructed by the addition of 0.76Å to the values of the covalent radii.[88-89] Numerous independent sets of radii have been proposed over the years from different approaches including potential energy curves, Slater-type orbitals, isodensity surfaces of the atomic wave function, and de Broglie wavelengths.[89] Bondi's initial set has been verified however by a number of statistical analyses of the contact distances in crystallographic data, and remains a valid representation of the van der Waal radii of the nonmetals despite a number of objections including the issue of the anisotrophy of atoms.[89] Different experimental approaches were required to obtain radii for metal species however now there are datasets that include transition metals, lathanides, and actinides.[87] Batsanov provides a review of many of the different compilations of van der Waal radii.[88] Emsley's handbook, 'The Elements'[90], is also often referenced because it was the first compilation of elemental data. The van der Waal radii reported in his book, 'The Elements', are taken from 'Lange's handbook of chemistry' and 'Physical data for inorganic chemists.'[90]

Recently, Truhlar and coworkers devised a set of radii based on relativistic coupled-cluster electronic structure calculations, that are compatible with Bondi's radii (which consists of only 28 of the 44 main-group elements), and therefore complete a dataset of van der Waal radii for the main-group elements.[89]

The latest compilation of experimentally derived radii is by Alvarez.[87] His radii are based on a distance distribution analysis of more than five million interatomic "non-bonded" distances.[87] This enables determination of a set of radii for most of the naturally occurring elements, however, again excluding the effects of anisotropy and multiple oxidation states.

Whilst the van der Waal radii form the basis of the molecular cavity structures, when the method of intersecting atomic spheres is employed, there are then differences regarding what is termed the solvent accessible surface (SAS). The original implementation of COSMO in MOPAC defines the SAS as follows: The solvent excluded surface (SEC) radii, R_A, is calculated from the van der Waals radii, R_A^{vdw}, plus, R^{solv}, which an effective radius for the solvent. However as the effective charges responsible for the dielectric screening are not located at the centre of the solvent molecules, another distance, δ^{SC} is specified. This distance is the range of 0.5 Å to R^{solv}, and in this study it is found empirically to be R^{solv} and is set to 1 Å. The minimum distance to a solute atom is $R_A^* = R_A - \delta^{SC}$. Therefore the distances R_A^*, which are just the van der Waal radii, are used to construct the SAS (Figure 6.4.1.1).[17]

The current COSab in the official release of GAMESS relies on radii shown to be appropriate for MP2 calculations by Baldridge & Jonas,[56] and the molecular shaped cavity is an extended algorithm of Klamt, where in addition to the above construction, additional features are added to account for crevices and T-shaped intersections in molecular surfaces.

Table 6.4.1-1 shows a summary of Pauling, Bondi, Emsley, Alvarez, and the optimized Klamt & Jonas radii, for a number of common elements. Also shown is the typically utilized 120% inflation of the radii. This radii inflation of 120% is an empirically derived expansion factor that has shown to be the optimal interaction radii for solute with solvent in the determination of free energies of solvation,[17] and is approximately the amount that the Klamt optimized radii are larger than the standard Bondi radii.[56]

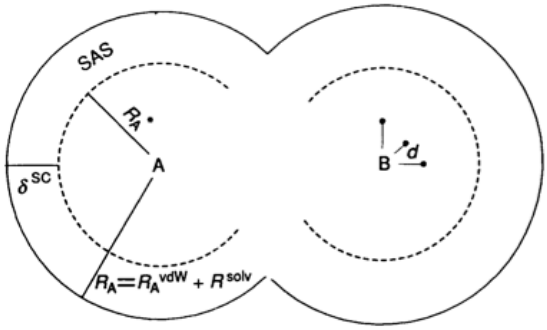


Figure 6.4.1.1 Construction of the SAS, takens from Klamt & Schüürman 1993[17] with permission from the Royal Society of Chemistry.

Table 6.4.1-1 Comparing the VDW radii of Pauling, Bondi, Emsley and Alvarez to the Klamt & Jonas optimized radii for a number of main atoms

Element		Pauling	Bondi[1]	Bondi + 20%	Emsley[2]	r_{vdW} Alvarez 2013	Alvarez + 20%	Klamt & Jonas optimized appropriate for MP2	% difference between Bondi & KlamtJonas radii
H	1	1.2	1.2	1.44	1.2	1.2	1.44	1.3	8.33
B	5	2.08			2.08	1.91	2.29	2.08	
C	6	1.7	1.7	2.04	1.85	1.77	2.12	2	17.65
N	7	1.5	1.55	1.86	1.54	1.66	1.99	1.83	18.06
O	8	1.4	1.52	1.82	1.4	1.5	1.80	1.72	13.16
F	9	1.35	1.47	1.76	1.35	1.46	1.75	1.72	17.01
Si	14	2	2.1	2.52	2	2.19	2.63	2.1	0.00
P	15	1.9	1.8	2.16	1.9	1.9	2.28	1.9	5.56
S	16	1.85	1.8	2.16	1.85	1.89	2.27	2.16	20.00
Cl	17	1.8	1.75	2.10	1.81	1.82	2.18	2.05	17.14
Br	35	1.95	1.85	2.22	1.95	1.86	2.23	-	
I	53	2.15	1.98	2.38	2.15	2.04	2.45	2.32	17.17

In an effort to test the effect of different vdW radii, several additional sets of radii were included in GAMESS, and two new key words were implemented to control user options, VDWRAD and VDWFAC. The VDWRAD option enables the user to select their choice of van der Waals radii from the options Bondi (VDWBON), Klamt & Jonas optimized radii (VDWKLM), Emsley (VDWEMS), and Alvarez (VDWALV). VDWFAC allows the user to select a radii expansion factor. The default radii are set to Bondi radii, and the default inflation factor is set to 1.2 (120%). Importantly, the user cannot add an expansion factor to the Klamt radii, as they are optimized radii and approximately 120% of the Bondi radii. Appendix D contains the necessary code additions for this added functionality.

GAMESS/COSab calculations were carried out with the Klamt, Bondi, Emsley and Alvarez radii, with an expansion factor of 1.2 (for all radii except the Klamt radii), comparing different OCE correction schemes (dblcav, dmulti and ocenon). The parameters of the cavity construction are: 1082 points for the basic grid, 162 segments on a complete sphere and a solvent radius of 1.3, and a dielectric permittivity of ϵ=78.4. Molecular structures were optimized with B97-D[36]/def2-TZVPPD[40] in both gas phase and solvent phase. Hessian calculations were not carried out. Tables 6.4.1-2, 6.4.1-3

and 6.4.1-4 report the ΔE_{solv} values for several test molecules including 17 neutral molecules, 4 anions and 5 cations. The results are consistent with the previous study by Baldridge and Klamt on the effect of OCE, without taking into consideration non-electrostatics and statistical thermodynamic considerations.[33] For neutral species the OCEs are typically under 1 kcal/mol, however this can still be a significant percentage of the overall ΔE_{solv} energy because the ΔE_{solv} energies of the neutral species are quite small. For anions the OCEs can be as large as 10 kcal/mol and because of the magnitude of the ΔE_{solv} of the anions the percentages reflect the importance of the OCE correction accurately. For cations the OCEs are typically smaller again and also represent a small percentage of the overall ΔE_{solv}. Importantly, differences are observed with the different radii. For example differences of up to 2 kcal/mol are observed between the computations with the Klamt radii and the 120% Alvarez radii for the neutral molecules, up to 4.3 kcal/mol differences are observed for the cations, and up to 6 kcal/mol differences are found for the anions.

Table 6.4.1-2 ΔE_{solv} of neutral systems with different VDW radii and OCE correction schemes

System	Exptl.	dblcav Klamt	dmulti Klamt	dblcav 1.2 Emsley	dmulti 1.2 Emsley	No OCE 1.2 Emsley	dblcav 1.2 Bondi	dmulti 1.2 Bondi	No OCE 1.2 Bondi	dblcav 1.2 Alvarez	dmulti 1.2 Alvarez	No OCE 1.2 Alvarez
HF	-5.6	-5.41	-5.32	-4.70	-4.84	-4.46	-4.34	-4.36	-4.23	-4.34	-4.36	-4.23
H_2O	-6.3	-6.96	-7.25	-6.09	-6.60	-5.61	-5.37	-5.56	-5.15	-5.37	-5.56	-5.15
NH_3	-4.2	-5.22	-5.54	-4.49	-4.85	-4.11	-4.41	-4.75	-4.06	-4.41	-4.75	-4.06
HC(O)H	-7	-4.46	-4.50	-3.86	-4.24	-3.59	-3.73	-3.81	-3.66	-3.73	-3.81	-3.66
C_2H_4	1.3	-1.63	-1.39	-0.97	-0.92	-0.97	-1.06	-1.15	-1.07	-1.06	-1.15	-1.07
CH_3OH	-5.1	-5.19	-5.41	-4.57	-5.01	-4.24	-4.04	-4.19	-3.90	-4.04	-4.19	-3.90
$HCONH_2$		-10.68	-10.89	-9.54	-10.19	-8.95	-8.68	-8.91	-8.45	-8.68	-8.91	-8.45
CH_2CCH_2		-1.85	-2.07	-1.41	-1.31	-1.38	-1.64	-1.64	-1.51	-1.64	-1.64	-1.51
CH_3NH_2	-4.6	-4.50	-4.83	-3.81	-4.23	-3.46	-3.77	-4.16	-3.44	-3.77	-4.16	-3.44
C_2H_6	1.8	-0.37	-0.23	-0.14	-0.11	-0.16	-0.07	-0.11	-0.09	-0.07	-0.11	-0.09
$CH_2CH_2CH_2$	0.8	-1.59	-1.57	-1.03	-0.96	-1.06	-1.21	-1.25	-1.21	-1.21	-1.25	-1.21
$CH_3CH_2CH_3$	2	-0.45	-0.29	-0.16	-0.15	-0.19	-0.07	-0.12	-0.09	-0.07	-0.12	-0.09
$(CH_3)_2NH$	-4.3	-3.53	-3.92	-2.92	-3.33	-2.69	-2.93	-3.31	-2.69	-2.93	-3.31	-2.69
$CH_3C(O)CH_3$	-3.9	-5.98	-6.25	-5.64	-6.03	-5.27	-4.90	-5.08	-4.82	-4.90	-5.08	-4.82
C_6H_6	-0.9	-2.89	-3.02	-1.31	-1.20	-1.33	-1.68	-1.65	-1.27	-1.68	-1.38	-1.61
$(CH_3)_3N$	-3.2	-2.42	-2.70	-2.01	-2.54	-1.94	-1.99	-2.28	-1.87	-1.99	-2.28	-1.86
$CH_3CH_2CH_2CH_3$	2.1	-0.80	-0.56	-0.34	-0.30	-0.37	-0.33	-0.35	-0.35	-0.33	-0.35	-0.35

Table 6.4.1-3 ΔE_{solv} of anion systems with different VDW radii and OCE correction schemes

System	Exptl.	dblcav Klamt	dmulti Klamt	dblcav 1.2 Emsley	dmulti 1.2 Emsley	No OCE 1.2 Emsley	dblcav 1.2 Bondi	dmulti 1.2 Bondi	No OCE 1.2 Bondi	dblcav 1.2 Alvarez	dmulti 1.2 Alvarez	No OCE 1.2 Alvarez
OH^-	-106	-95.68	-97.85	-97.31	-99.91	-86.45	-90.00	-91.76	-82.10	-90.00	-91.76	-82.10
NO_2^-	-72	-72.23	-73.11	-73.22	-74.28	-67.89	-69.50	-70.16	-65.81	-69.50	-70.16	-65.81
NH_2^-	-93	-91.07	-92.66	-89.44	-90.96	-80.98	-88.88	-90.30	-80.66	-88.88	-90.30	-80.66
$CH_3CO_2^-$	-77	-73.79	-74.96	-73.92	-75.76	-68.85	-69.78	-70.36	-66.22	-69.78	-70.36	-66.22

Table 6.4.1-4 ΔE_{solv} of cation systems with different VDW radii and OCE correction schemes

System	Exptl.	dblcav Klamt	dmulti Klamt	dblcav 1.2 Emsley	dmulti 1.2 Emsley	No OCE 1.2 Emsley	dblcav 1.2 Bondi	dmulti 1.2 Bondi	No OCE 1.2 Bondi	dblcav 1.2 Alvarez	dmulti 1.2 Alvarez	No OCE 1.2 Alvarez
CH_3^+		-78.68	-78.65	-71.95	-71.96	-72.36	-75.74	-75.73	-76.31	-75.74	-75.73	-76.31
NH_4^+	-77	-81.72	-81.56		-77.67	-78.40	-77.35	-77.30	-77.99	-77.35	-77.30	-77.99
$(CH_3)_2NH_2^+$	-61	-64.63	-64.63	-60.55	-60.43	-61.52	-61.08	-61.21	-62.28	-61.08	-61.21	-62.28
$(CH_3)_3NH^+$	-57	-58.53	-58.52	-54.93	-54.89	-55.97	-55.62	-55.70	-56.97	-55.62	-55.70	-56.97
$(CH_3)_3O^+$		-56.59	-56.49	-53.35	-53.31	-54.36	-54.33	-54.35	-55.60	-54.33	-54.35	-55.60

6.4.2 Isodensity contour surfaces

Isodensity surfaces are of interest for several reasons. Firstly, as discussed at the beginning of this chapter, the distance dependent dielectric scheme proposed requires knowledge of how the electron density changes at distance r from the vdW cavity points. In fact, it may be that an on-the-fly calculation would be optimal for such an implementation. Secondly, the isodensity surface may provide a more rigorous boundary for a double cavity OCE strategy. Finally, as the literature is lacking an extensive examination optimal isodensity contour surfaces across varying system types, it is of interest to conduct such a study, which will further aid the implementation of the these ideas.

A straightforward approach was conceived of for the implementation of the isodensity surface into COSab, involving using the existing cavity construction surface routine based on vdW radii, as the starting point. In the new algorithm, the initial cavity is constructed as previously done, but then the electron density information is obtained at each segment of this cavity. From this point, a new routine is incorporated that enables expansion or contraction of cavity points until the desired isodensity contour is reached (within a given tolerance). The advantage of the isodensity method is that it makes use of pre-existing cavity routine that has been optimized to avoid numerical instabilities and problems with crevices or overlapping spheres.[3]

The new cavity construction options offer two possible strategies that are equally worth exploring. The first is an isodensity surface implementation for the primary surface, coupled to the DMULTI OCE method, or, without any OCE correction. The second option particularly interesting to the COSab algorithm would be to use a vdW cavity for the primary cavity, and an isodensity cavity as the secondary cavity for the double cavity OCE correction. The algorithmic flow accommodating both of these strategies to the isodensity cavity subroutine as implemented in the COSCAV subroutine, is shown in Figure 6.4.2.1.

To enable these capabilities in COSab, two new key words were created; ISOCAV and COSDEN. ISOCAV is a logical (true/false) that switches on the isodensity cavity routine, and COSDEN specifies the isodensity contour value.

The subroutine for creating the isodensity surface is called ZROFLX, and is invoked if ISOCAV is set to true. ZROFLX takes the already determined grid points on the cavity surface and calls a routine already available in GAMESS to calculate the electron density at given points. Once the electron density at the surface points is known, then the cavity is either expanded or contracted to reach the set COSDEN value plus/minus 15%.

```
 3700     CONTINUE
 3800  ILIPA=ILIPA+NIPA(IA)
      COSVOL=COSVOL/3
C HERE IS THE ENDIF ASSOCIATED WITH THE
C IOLDCV VARIABLE CONTROLING CAVITY CLOSURE
C      END IF                                          <--
      COSVOL=COSVOL/(TOANGS*TOANGS*TOANGS)
C REPLACE NPS BY NPSPHER
      IF(OUTCHG.EQ.DBLCAV) THEN
      DO 449 IPS=1,NPSPHER
          JPS=NPS+IPS
          I=IATSP(IPS)
          IATSP(JPS)=I
          RI=(SRAD(I)+COSRAD)+(ROUTF-1)*COSRAD
          NFA(JPS)=NFA(IPS)
          NSETF(JPS)=NSETF(IPS)
          DO 448 IX=1,3
             XSP(IX,JPS)=COORD(IX,I)+
     $                   (XSP(IX,IPS)-COORD(IX,I))*RI/SRAD(I)
 448       CONTINUE
          FL(JPS)=FL(IPS)*(RI/SRAD(I))**2
 449   CONTINUE
       ENDIF
      IF (NPS.GT.MXABC) THEN
       IF(MASWRK) THEN
       WRITE(IW,*) 'NPS IS GREATER THAN MXABC (4000), NPS=',NPS
       WRITE(IW,*) 'THE CAVITY IS TOO LARGE TO BE HANDLED, REFER'
       WRITE(IW,*) 'TO THE MANUAL TO INCREASE GAMESS LIMITATIONS'
       END IF
       RETURN
      ENDIF                                            <--
      NPSD= NPS + NPSPHER
C SAVE XSP TO COSURF
      DO L=1,NPSD
          COSURF(1,L) = XSP(1,L)
          COSURF(2,L) = XSP(2,L)
          COSURF(3,L) = XSP(3,L)
      ENDDO
```

Figure 6.4.2.1 Algorithmic flow of the COSCAV subroutine, indicating where the calls to the isodensity subroutine, ZROFLX can be made.

The double cavity implementation already offered an algorithm for expanding the VDW cavity surface in order to form the double cavity. As an aside, when working through this formula, it was observed that this expression resulted in an unintuitive expansion of the cavity, in that the cavity pushed atoms with small radii (e.g. Hydrogen) out further than atoms with larger radii (e.g. Oxygen). Expanding the formula of the new double cavity points by substituting in the formula for RI and the applying the ROUTF of 0.85 now results in the following,

XSP(IX,JPS)=COORD(IX,I)+ (XSP(IX,IPS)-
COORD(IX,I))*(SRAD(I)+0.85*COSRAD)/SRAD(I)

where COORD(IX,I) are the atomic starting coordinates (along the x, y, and z axes), therefore making, XSP(IX,IPS)-COORD(IX,I) the distance between the surface points and the related atomic coordinate. This distance is multiplied by SRAD(I) + 0.15*COSRAD, which essentially extends past the existing cavity by 15% of COSRAD, but then it is divided by SRAD(I). SRAD(I) is the given vdW radii for each atom type. Therefore, dividing by SRAD pushes out by a larger amount for Hydrogen than would be for heavy elements, such as Oxygen. Whilst the equation was found by fitting to experiment, this remains a rather unsatisfying result.

For the isodensity cavity capability, a more straightforward expansion formula was developed that simply takes the distance between the atom and the cavity point, and multiplies it by a pushing factor (in and out for contraction, expansion respectively),

SURPTS(IX,IPOINT)=COORD(IX,I)+ (SURPTS(IX,IPOINT)-
COORD(IX,I))*PUSHIN

When the convergence range is reached, the routine moves on to the next point and so on until all the cavity points lie within the selected range of electron density value. Then, this resulting set of points is fed back into to the COSCAV routine.

The $ELDENS group is a dollar group, available in the official release of GAMESS, that controls the electron density calculation. The keyword WHERE designates where the electron density is to be calculated (e.g. at the center of mass, at each nuclei, at a set of points, or on a grid). If WHERE=POINTS is specified, a set of points is required in a separate dollar group, $POINTS, specification. For the new isodensity feature,

this capability was exploited for calculation of the electron density at the surface points. As currently the input is expected in the input file, a few alterations to the source files were required to have the input fed on the fly from the calculation. Two source files required changes, prpel.src and prplib.src. The subroutine prplib.src prepares the coordinates of the next point at which the electron density is to be calculated and sends the coordinate to the subroutine, ELDENC in prpel.src. Firstly, the number of maximum points permitted was expanded from MXPTPT=100 to MXPTPT=1000. Secondly, the ISODAT common block was added, and thirdly, a common block was added for the property data, ELPROP. ISODAT is a newly defined common block to transfer the electron density information between the relevant subroutines and any other newly created data that may be important for the isodensity cavity. In prpel.src the following modifications were made to the ELDENC subroutine:

1) The COSDAT and ISODAT common blocks are added.
2) At the end of the subroutine before 'GO TO 210', the following is added:

```
IF(ISOCAV) THEN
        CAVDEN(IPOINT) = EDENS
ENDIF
```

This construct specifies storage of the electron densities as they are calculated in an array, CAVDEN, which is identified in the ISODAT common block.

Figures 6.4.2.2 and 6.4.2.3 provide graphical representations of the cavity charge locations for water (Figure 6.4.2.2) and propane (Figure 6.4.2.3) with various cavity representations achieved with this isodensity algorithm. The cavities appear quite reasonable expansions of the vdW cavity, indicating algorithmic stability and validating the scheme for identification of the isodensity surface points.

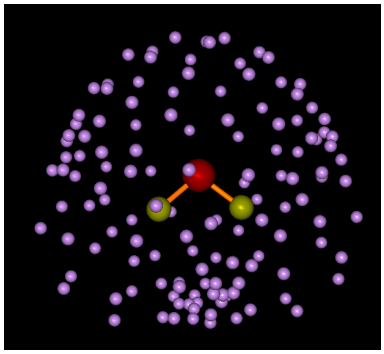
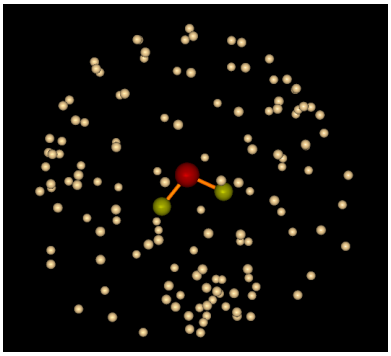
(a)  (b) 

Figure 6.4.2.2 Isodensity cavities for water at (a) 0.01 a.u. and (b) 0.001 a.u.

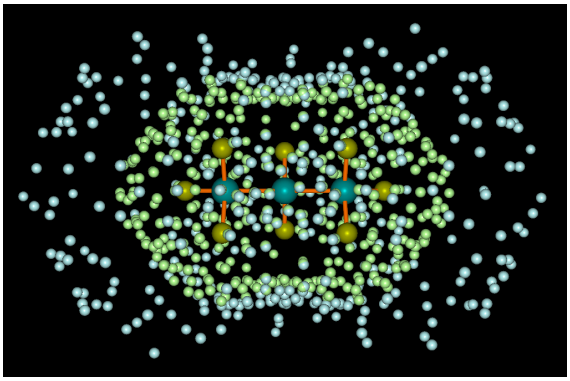


Figure 6.4.2.3 Propane cavities; green: vdw cavity; blue: isodensity cavity at 0.01 a.u.

6.4.3 General radial dependence

Radial dependent analyses are useful in demonstrating the validity of a continuum model. Theory predicts an 1/r dependence for symmetric ions (monopoles), approximately an $1/r^3$ dependence for dipolar neutrals, $1/r^5$, for quadropolar compounds, and $1/r^7$ for octopolar compounds. While the isodensity studies specified an isodensity contour, rather than a radii choice, it is easy to find the distance at each isodensity contour for the monovalent atomic ions to test the models.

The monoatomic test set consisting of K^+, Na^+, F^- and Cl^- turned out to be very insightful in the investigation of completeness of the methodology implementation. Inconsistent results emerged in calculations at the same isodensity contour with different vdW radii starting points. This indicated that terms dependent on the initial vdW cavity were still being exploited at the point where calls to the isodensity routine were being made. Examination of the code revealed that, in fact, the

interaction matrix elements belonging to the A matrix (i.e., solvent/solvent interactions) are calculated in the section that follows the cavity construction. These matrix elements (equation 6.4.3-1) require both the positions, r, and the areas, S, of the segment patches of the cavity surface. The isodensity routine returns the new positions of the center point of each surface patch, however, as the surface is either contracted or expanded, the original surface patches are obviously no longer of the same surface area, and furthermore as the expansion or contraction is not uniform, a simple mapping of the patches is not possible.

$$a_{\mu v} = \frac{1}{|S_\mu||S_v|}\int_{S_\mu}\int_{S_v} \|r - r'\|^{-1} d^2r' d^2r \qquad (6.4.3\text{-}1)$$

As such, an additional surface construction algorithm will be required to determine the segment areas from the new isodensity surface. A surface construction algorithm such as the Voronoi tessellation scheme, built from the new isodensity points would generate Voronoi cells that can be used to obtain the surface areas and hence enable the isodensity procedure. Tessellation schemes are currently under investigation.

To provide a proof of principle that the isodensity routine would work with new surface patch information, a simple test was conducted whereby the isodensity calculations were run with various isodensity contour values in order to obtain the distance that the isodensity contour lies from the atom center, and hence define a new radii to represent each isodensity contour. This is of course only possible for the calculations of atoms because the contraction or expansion to the isodensity cavity is uniform. Once these distances, $R_{isodensity}$, were obtained, a VDWFAC was obtained via equation 6.4.3-2.

$$VDWFAC = \frac{R_{isodensity}}{VDWRAD} \qquad (6.4.3\text{-}2)$$

This VDWFAC could then be applied, using whichever VDWRAD were selected to obtain the VDWFAC, with the DMULTI OCE method, to obtain the correct energetics at the radii corresponding to the isodensity contour. VDWFACs were found for

isodensity contours of 0.01, 0.008, 0.006, 0.004, 0.002 and 0.001 (±15%) for K^+, Na^+, F^- and Cl^- and compared to the Klamt, Emsley, Bondi and Alvarez radii (Figures 6.4.3.1 – 6.4.3.4 and Tables 6.4.3-2 – 6.4.3-5). General trends were sought rather than exact values primarily because a range of experimental values for the ΔG_{solv} are reported in the literature for each of the monovalent atomic ions (Table 6.4.3-1). Also, the values considered are only the E_{solv} values, with no consideration of the statistical thermodynamic functions.

The Klamt radii for sodium and potassium are identical at 2.31 Å. We note that in the case of K^+, the radii closest to yielding the experimental ΔG_{solv} are the Klamt radii. In the case of Na^+, this is not the case, and the shorter radii producer a closer prediction. For the anions, F^- and Cl^-, the Klamt radii are clearly the best choice to achieve the solvation energy values. In fact, it has been previously shown for ions that a radius of 1.10 – 1.15 times the van der Waal radii is recommended, as opposed to the 1.2 in the case of neutral molecules.[91]

As evident from the experimental data, negative ions are more stably solvated compared to positive ions. Hummer et al. suggest that this is due to the structural asymmetry of water that allows the hydrogen atoms to "penetrate the ionic van der Waals shell," however, "the oxygen atom is better protected."[92] This is likely to have implications regarding the missing short-range interactions and may be explain the electron density surface results. For Na^+, the Klamt radii sit at an electron density of 0.001400 a.u.. We found that, whilst a 0.01 a.u. electron density contour reproduces the experimental results for Na^+ fairly well, in the case of K^+, such a contour value is too small to obtain a small enough radius to produce hydration energies within the experimental range. The electron density value for the Klamt surface of potassium is 0.019 a.u. This value is significantly higher than the tested range of isodensity contours. For the anions, F^- and Cl^-, the isodensity contour within the range 0.01 – 0.001 are significantly too low to get anywhere near the experimental range of hydration energies. The electron density for the Klamt surface is 0.052072 a.u. for fluorine and is 0.061472 a.u. for chlorine. These results are consistent with Chipman's observation that unless "unphysically small cavities" are employed for the calculation of anion solvation energies in water, they are significantly underestimated by

electronic structure calculations.[70] He further concludes that this effect is likely due to the strong first solvation shell interactions.[70] For charged systems generally it has been shown that, because of the importance of the short-range interactions, the screening by a polarizable continuum is only appropriate beyond 7 Å.[29] Therefore, one would not want to place too much weight on the results from these isodensity studies. However, the demonstrated radial depends validates the isodensity algorithm.

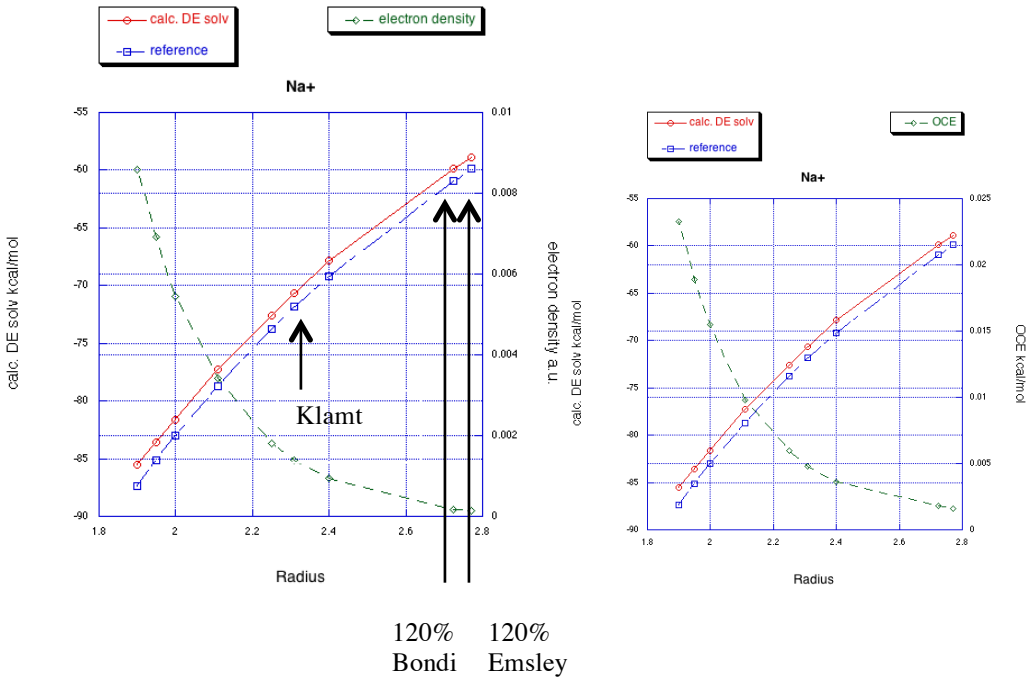


Figure 6.4.3.1 Effect of radii on electrostatic solvation energy for Na^+ and (a) relationship between electron density and radii and (b) effect of radii on outlying charge

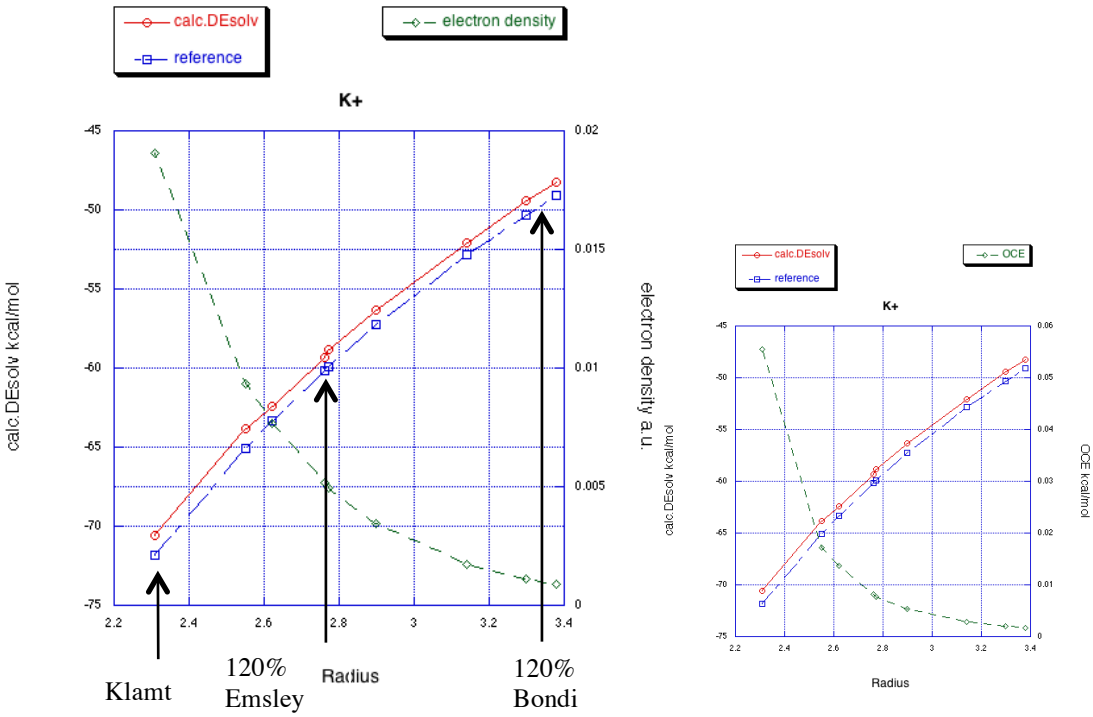


Figure 6.4.3.2 Effect of radii on electrostatic solvation energy for K^+ and (a) relationship between electron density and radii and (b) effect of radii on outlying charge

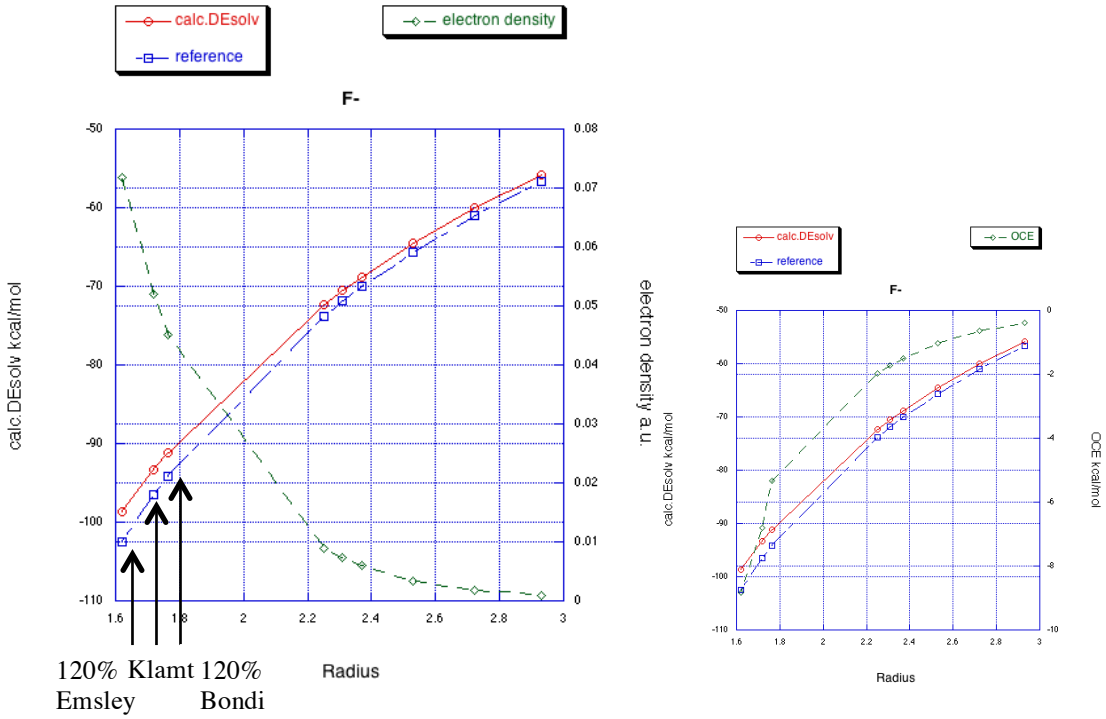


Figure 6.4.3.3 Effect of radii on electrostatic solvation energy for F^- and (a) relationship between electron density and radii and (b) effect of radii on outlying charge

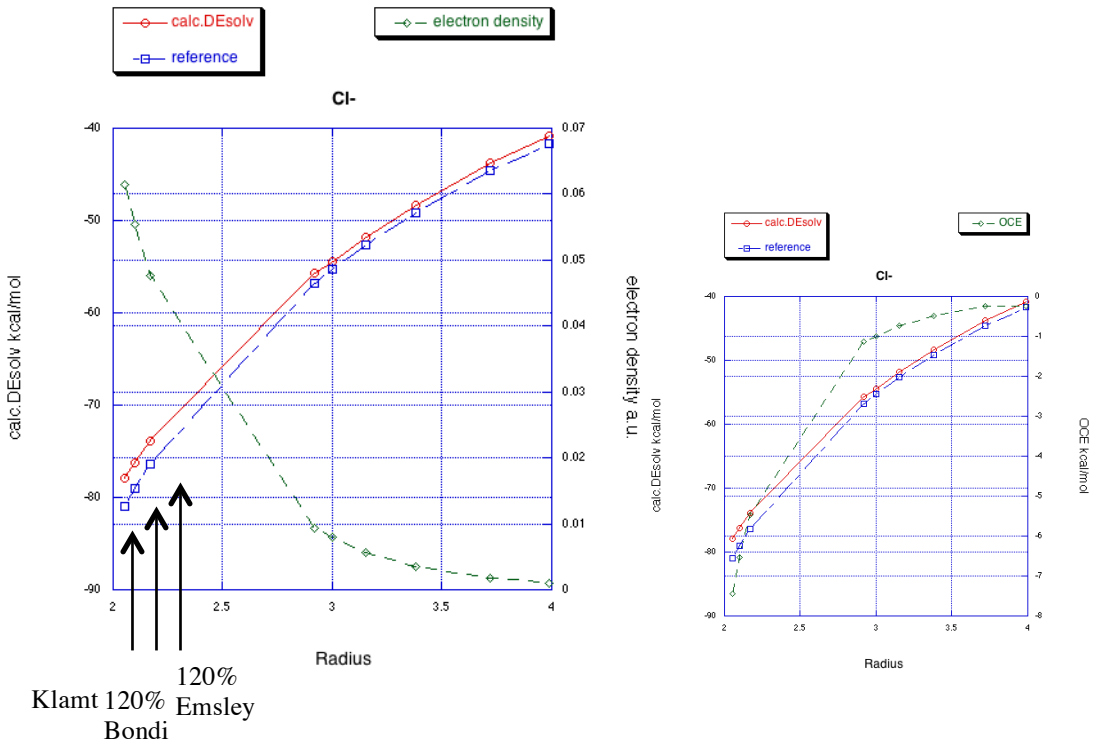


Figure 6.4.3.4 Effect of radii on electrostatic solvation energy for Cl^- and (a) relationship between electron density and radii and (b) effect of radii on outlying charge

Table 6.4.3-1 Experimental free energy of ionic hydration (kJ/mol) reported in J. Phys. Chem. Vol 100 No. 4 1996

	Marcus	Friedman & Krishnan	Conway
Na^+	-87.2	-88.6	-88.9
K^+	-70.5	-71.2	-71.2
F^-	-111.1	-94.1	-105.3
Cl^-	-81.2	-66.2	-77.4

Table 6.4.3-2 Isodensity surface results for Na^+

Isodensity surface (a.u.)	0.008574	0.006919	0.005435	0.003424	0.001795	0.000942
OCE (kcal/mol)	0.0233	0.0189	0.0155	0.0098	0.0060	0.0036
Radius	1.90	1.95	2.00	2.11	2.25	2.40
VDWFAC	0.84	0.86	0.88	0.93	0.99	1.06
SP-GP kcal/mol	-85.56	-83.57	-81.67	-77.28	-72.60	-67.80

Table 6.4.3-3 Isodensity surface results for K^+

Isodensity surface (a.u.)	0.009356	0.007688	0.005148	0.003439	0.001752	0.000879
OCE (kcal/mol)	0.0172	0.0137	0.0082	0.0053	0.0028	0.0016
Radius	2.55	2.62	2.76	2.90	3.14	3.38
VDWFAC	0.93	0.95	1.00	1.05	1.14	1.23
SP-GP kcal/mol	-63.79	-62.45	-59.33	-56.50	-52.04	-48.23

Table 4.2.3.7-4 Isodensity surface results for F^-

Isodensity surface (a.u.)	0.008937	0.00729	0.005944	0.003438	0.001771	0.000875
Outlying charge effect (kcal/mol)	-1.97	-1.73	-1.51	-1.03	-0.65	-0.38
Radius	2.25	2.31	2.37	2.53	2.72	2.93
VDWFAC	1.53	1.57	1.61	1.72	1.85	1.99
SP-GP kcal/mol	-72.35	-70.54	-68.82	-64.47	-59.97	-55.76

Table 6.4.3-5 Isodensity surface results for Cl^-

Isodensity surface (a.u.)	0.009394	0.00794	0.005729	0.00346	0.001701	0.000976
Outlying charge effect (kcal/mol)	-1.13	-0.99	-0.74	-0.48	-0.24	-0.14
Radius	2.92	3.00	3.15	3.38	3.72	3.99
VDWFAC	1.67	1.71	1.80	1.93	2.13	2.28
SP-GP kcal/mol	-55.70	-54.42	-51.74	-48.28	-43.76	-40.89

6.5 Outlook

This final chapter has outlined a number of the considerations required in the implementation of a distance dependent dielectric function that depends on the electron density of the solute system, and therefore has paved the way towards a more accurate solvent model. To this goal, options to increase the user control and flexibility over the vdW cavity construction routine were added, the distributed multipole routine was parallelized, an option to switch off OCE correction entirely was added and an isodensity cavity routine has been developed. A flowchart for the COSab implementation within the GAMESS software, with these additional capabilities that involve a modification of the algorithmic flow, is illustrated in Figure 6.5.1.

With the information now obtainable from the isodensity cavity scheme, a small thought experiment was carried out to demonstrate the potential impact of these developments. The surface charge coordinates for the vdW cavity (Klamt radii) surface, the 0.01 a.u. isodensity surface and the 0.001 a.u. isodensity surface were obtained for acetic acid. These were plotted and the images were overlaid along with the continuum cluster of acetic acid with the $S_C[Q_{a1}Q_{br}Q_{a2}]$ solvent network (Figure 6.5.2). The bounds of the explicit waters lined up nicely with the 0.001 cavity surface. This may indicate that the 0.001 a.u. surface provides a good estimate of where the bulk continuum properties can be applied, providing further motivation for exploring a three-layer solvent model, where the dielectric is stretched out to a particular isodensity value. Further investigations are however required to verify this hypothesis.

The inclusion of a tessellation scheme and the surface area of the segment patches should complete the isodensity cavity scheme, and allow testing of the solvation energies at various isodensity values. This implementation would then be used to start exploring different functions to stretch out the dielectric function from the cavity to the bulk dielectric value, in a distance dependent dielectric scheme. Further work is also required to consider how to introduce the explicit solvent interactions implicitly or return to some of the ideas tried in Chapter 4.4.

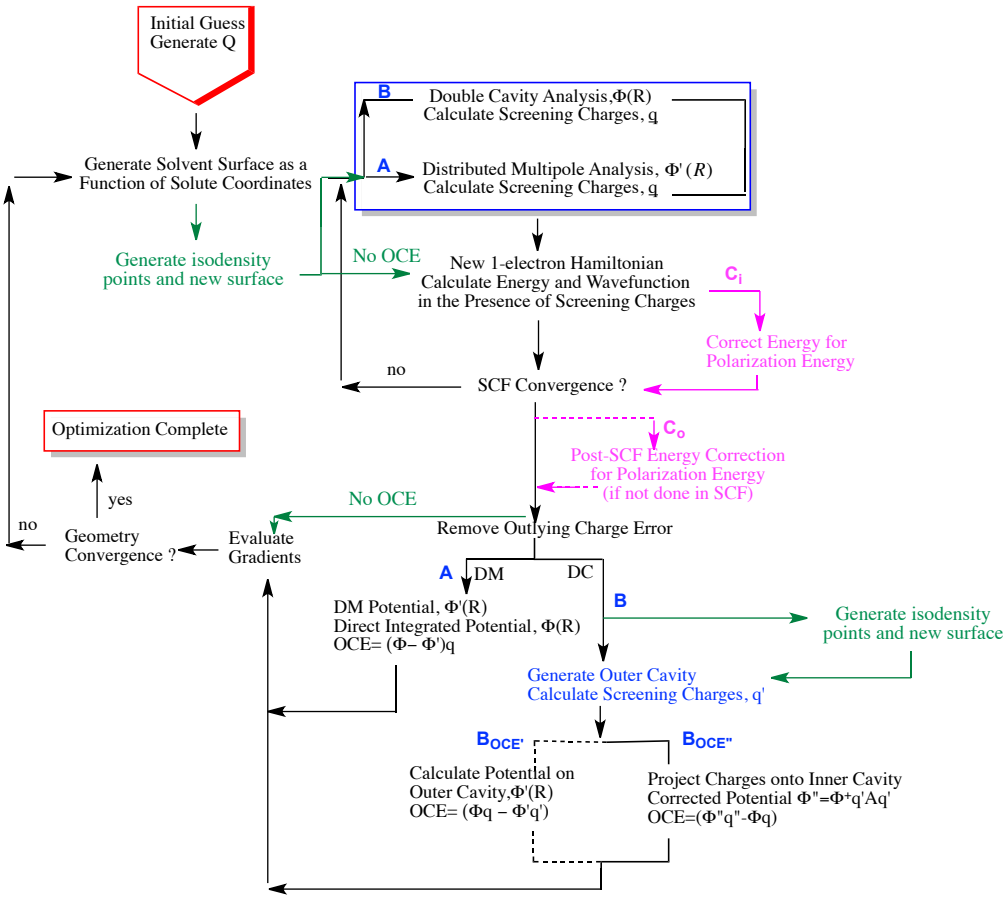


Figure 6.5.1 Flow diagram for COSab with new modifications (green) within the Hartree-Fock-SCF procedure

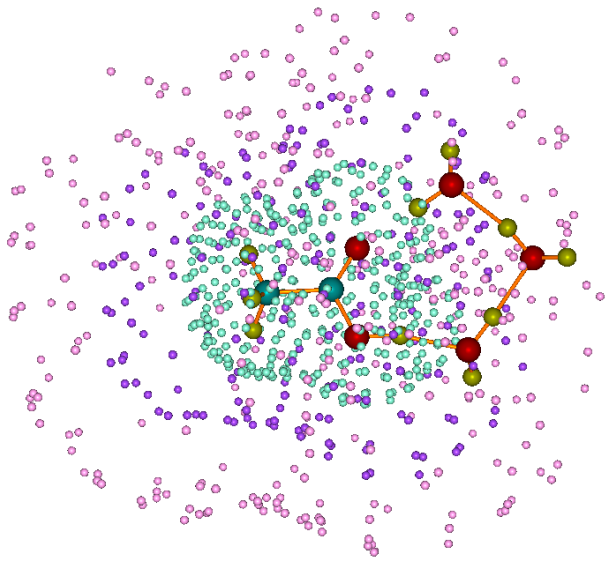
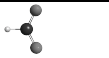
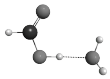
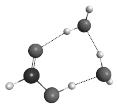
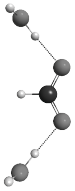
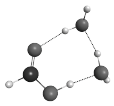
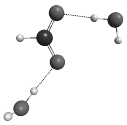
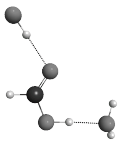
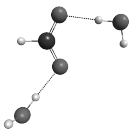
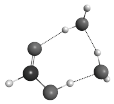
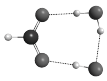
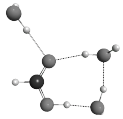
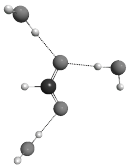
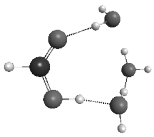
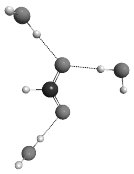
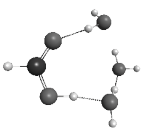
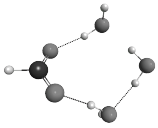
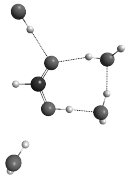
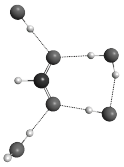


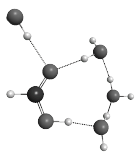
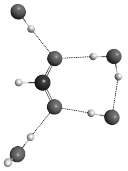
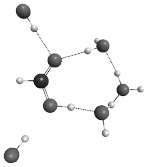
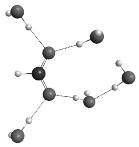
Figure 6.5.2 Overlaid cavity surfaces for acetic acid (pink cavity: 0.001 a.u. isodensity cavity; purple cavity: 0.01 a.u. isodensity cavity; green cavity: vdW (Klamt) cavity) over the acetic acid $S_C[Q_{a1}Q_{br}Q_{a2}]$ cluster.

Appendix A

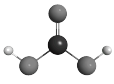
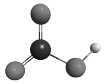
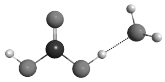
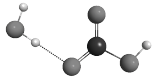
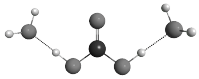
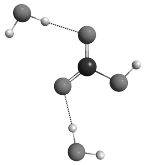
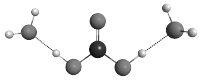
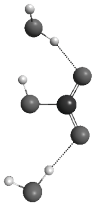
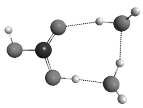
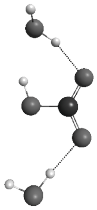
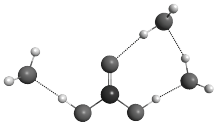
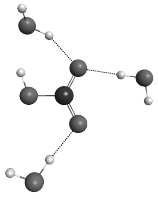
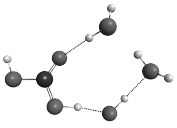
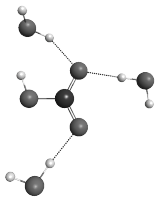
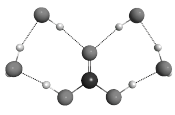
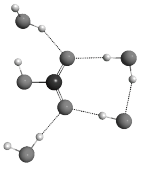
DSES-CC clusters

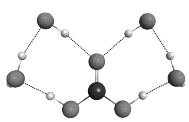
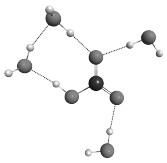
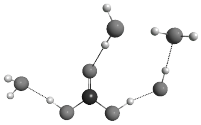
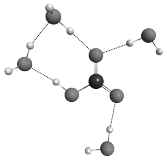
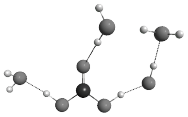
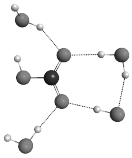
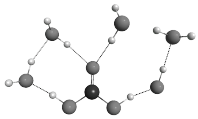
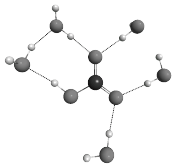
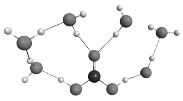
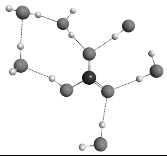
Formic acid (exptl. pK_a = 3.77)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			2.89	0.88
1			3.74	0.03
2			3.22	0.55
2			3.59	0.18
2			2.34	1.43
2			3.48	0.29
3			3.04	0.73
3			3.09	0.68
3			3.53	0.24
4			0.58	3.19

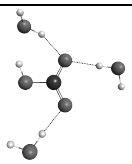
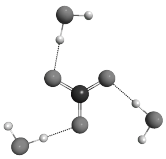
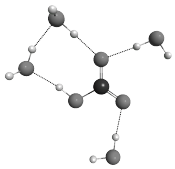
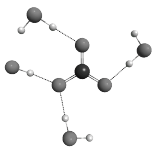
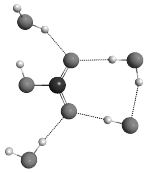
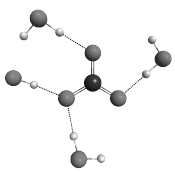
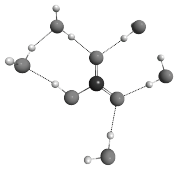
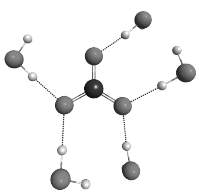
4			1.95	1.82
5			0.74	3.03

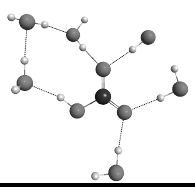
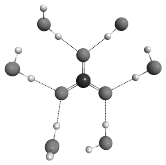
Carbonic acid (exptl. $pK_a = 3.58$)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			1.47	2.11
1			3.18	0.40
2			3.16	0.42
2			3.02	0.56
2			2.17	1.41
3			3.61	-0.03
3			2.23	1.35
4			3.85	-0.27

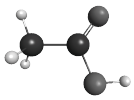
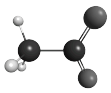
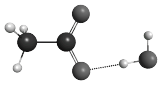
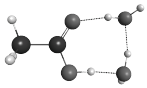
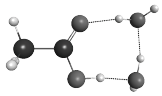
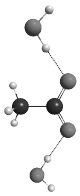
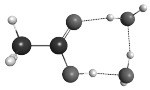
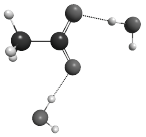
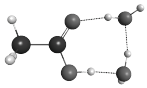
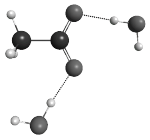
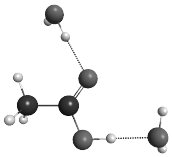
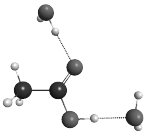
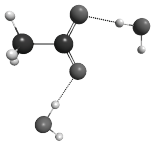
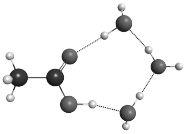
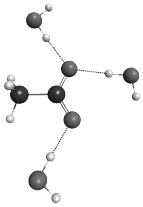
4			3.56	0.02
4			3.05	0.53
4			3.35	0.23
5			2.84	0.74
6			2.44	1.14

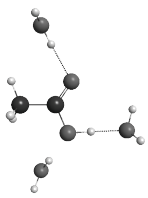
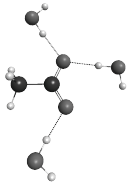
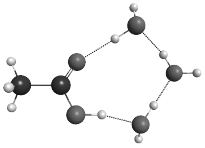
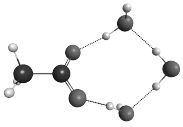
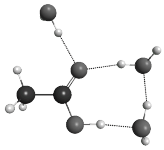
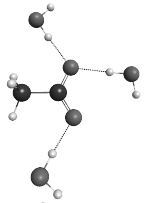
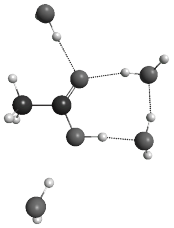
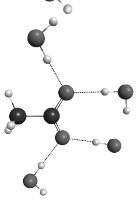
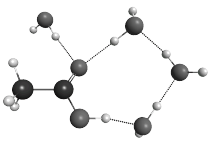
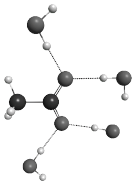
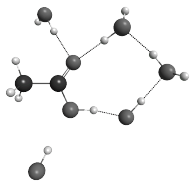
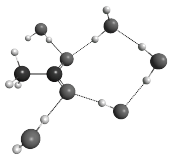
Carbonate (exptl. $pK_a^2 = 10.6$)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3			13.93	-3.33
4			13.19	-2.59
4			12.90	-2.30
5			10.95	-0.35

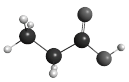
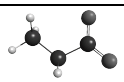
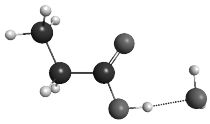
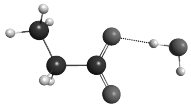
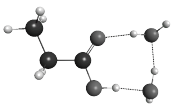
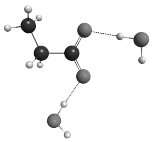
6			8.98	1.62

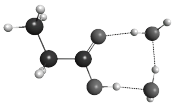
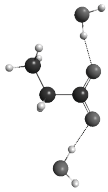
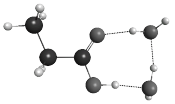
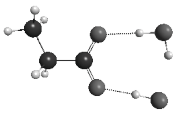
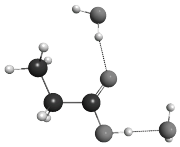
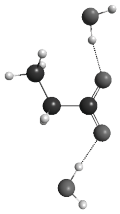
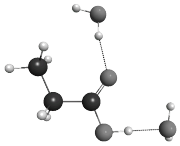
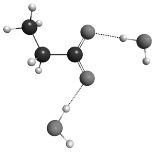
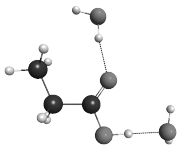
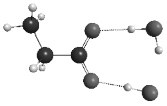
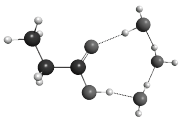
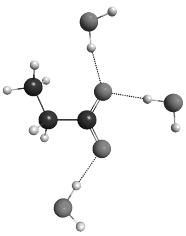
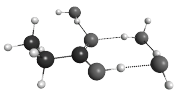
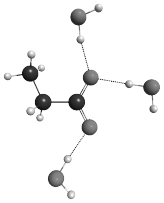
Acetic acid (exptl. $pK_a = 4.76$)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.20	-1.44
1			6.23	-1.47
2			5.30	-0.54
2			4.73	0.03
2			4.94	-0.18
2			5.30	-0.54
2			4.10	0.66
2			3.74	1.02
3			4.58	0.18

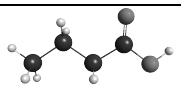
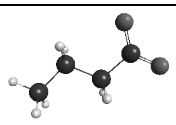
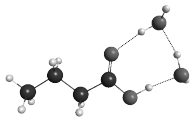
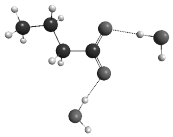
3			1.63	3.13
3			5.41	-0.65
3			4.39	0.37
4			2.34	2.42
4			3.55	1.21
5			2.51	2.25

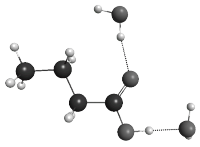
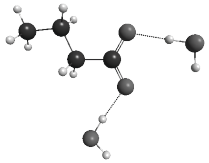
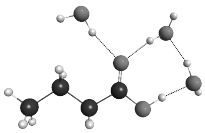
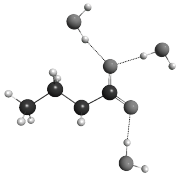
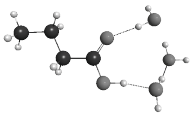
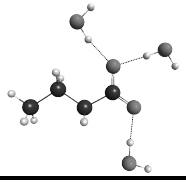
Propanoic acid (exptl. $pK_a = 4.86$)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.74	-1.88
1			6.59	-1.73
2			5.90	-1.04

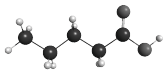
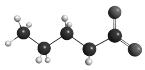
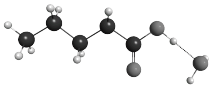
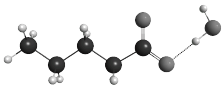
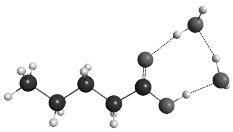
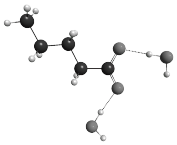
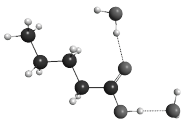
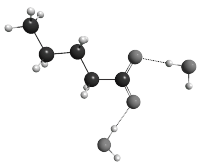
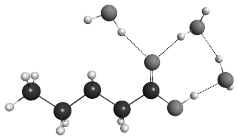
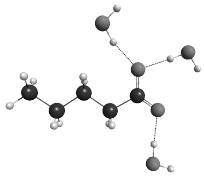
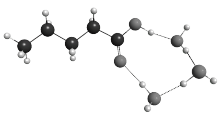
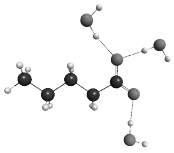
2			5.66	-0.80
2			5.85	-0.99
2			4.68	0.18
2			4.92	-0.06
2			4.87	-0.01
3			5.65	-0.79
3			5.26	-0.40

Butanoic acid (exptl. $pK_a = 4.83$)

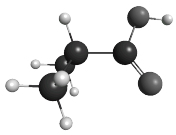
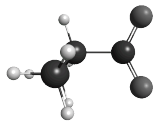
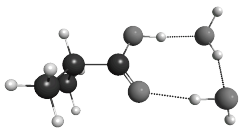
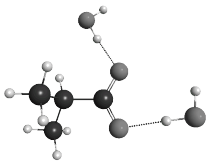
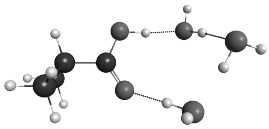
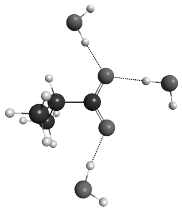
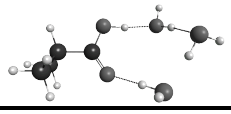
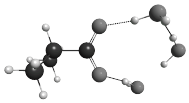
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.76	-1.93
2			5.91	-1.08

2			4.45	0.38
3			4.94	-0.11
3			5.24	-0.41

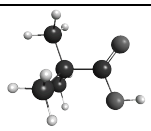
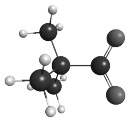
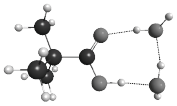
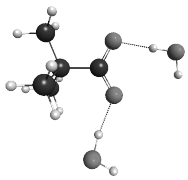
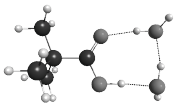
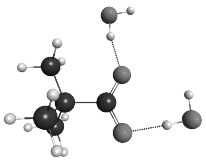
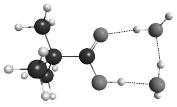
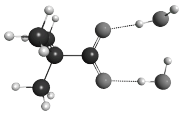
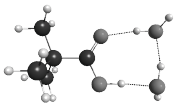
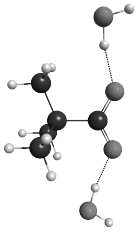
Pentanoic acid (exptl. $pK_a = 4.84$)

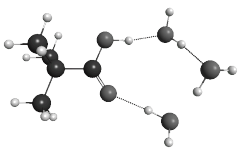
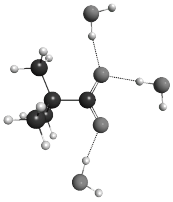
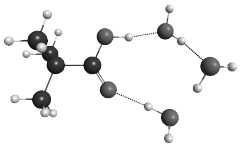
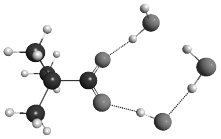
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.62	-1.78
1			6.51	-1.67
2			5.84	-1.01
2			4.81	0.03
3			4.99	-0.15
3			5.26	-0.42

Isobutyric acid (exptl. pK_a = 4.88)

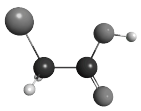
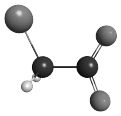
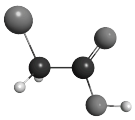
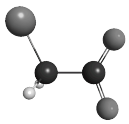
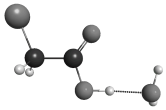
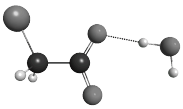
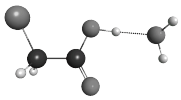
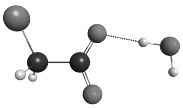
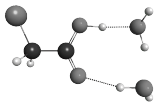
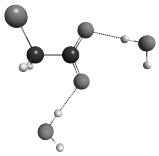
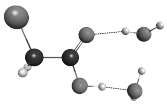
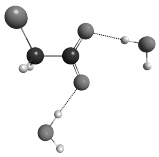
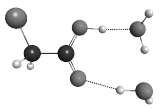
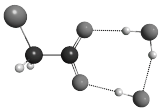
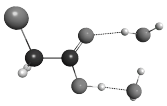
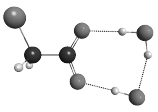
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.59	-1.71
2			5.91	-1.03
3			5.11	-0.23
3			5.93	-1.05

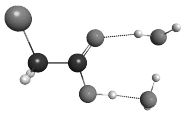
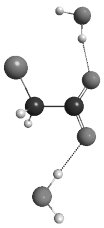
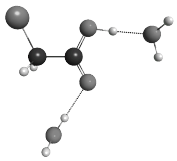
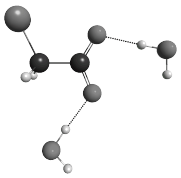
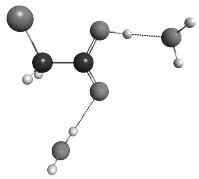
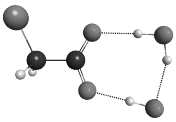
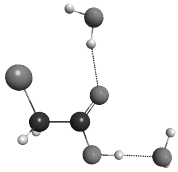
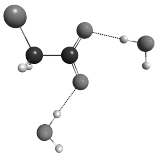
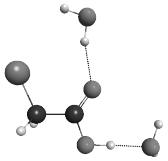
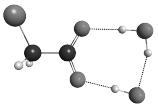
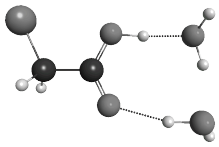
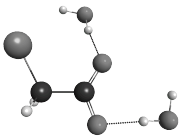
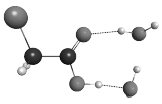
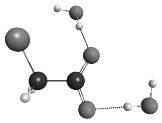
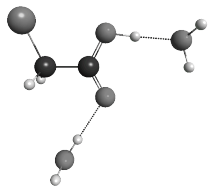
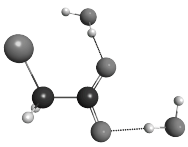
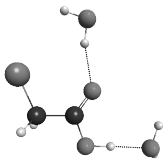
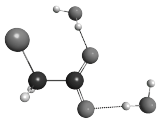
Trimethylacetic (Pivalic) acid (exptl. pK_a = 5.03)

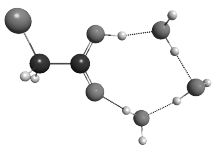
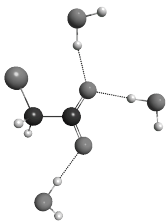
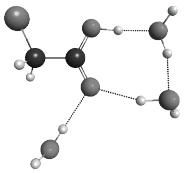
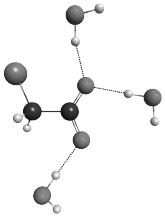
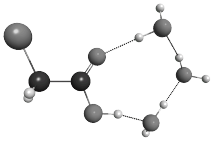
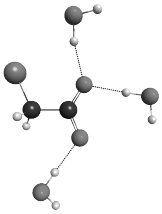
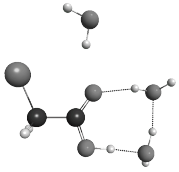
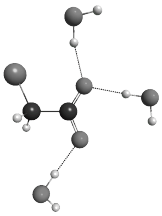
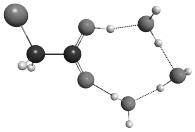
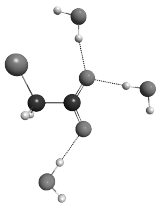
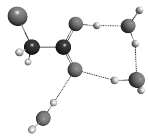
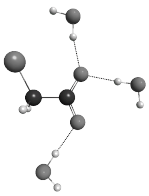
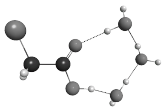
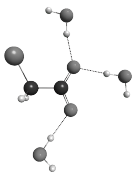
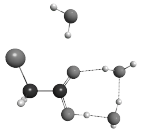
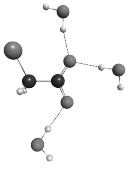
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			7.05	-2.02
2			6.28	-1.25
2			6.79	-1.76
2			6.00	-0.97
2			6.63	-1.60

3			5.37	-0.34
3			6.19	-1.16

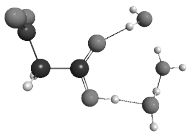
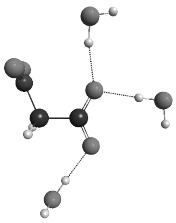
Chloroacetic acid (exptl. $pK_a = 2.81$)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			1.82	0.99
0			1.59	1.22
1			2.46	0.35
1			2.13	0.68
2			2.23	0.58
2			2.61	0.20
2			2.02	0.79
2			2.40	0.41

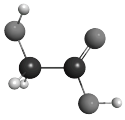
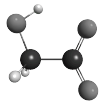
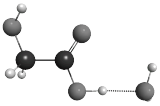
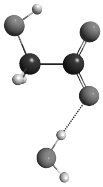
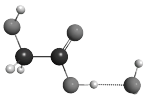
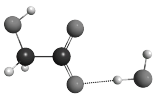
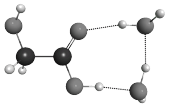
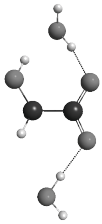
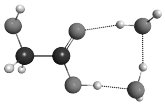
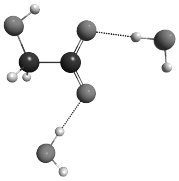
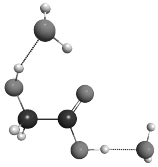
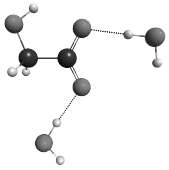
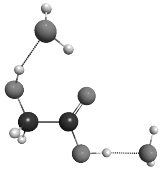
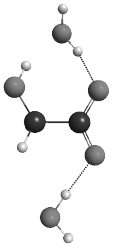
2			2.44	0.37
2			1.22	1.59
2			1.01	1.80
2			1.04	1.77
2			0.83	1.98
2			2.04	0.77
2			2.42	0.39
2			1.03	1.78
2			0.85	1.96

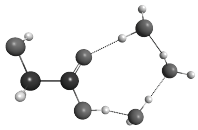
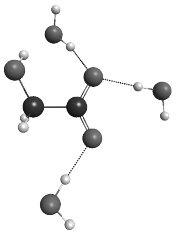
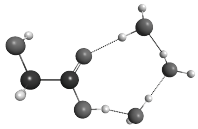
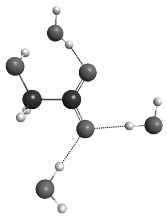
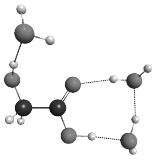
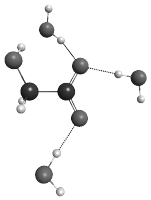
3			2.30	0.51
3			1.72	1.09
3			2.64	0.17
3			2.17	0.64
3			1.90	0.91
3			1.32	1.49
3			2.25	0.56
3			1.77	1.04

Nitroacetic acid (exptl. pK_a = 1.32)

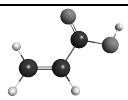
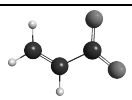
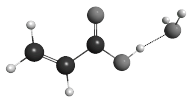
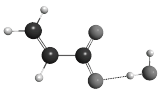
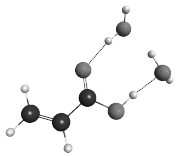
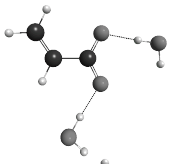
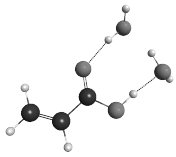
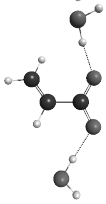
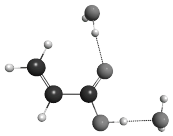
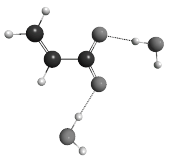
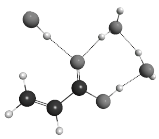
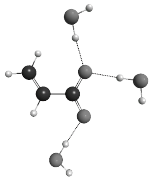
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3			1.49	-0.17

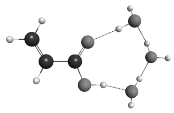
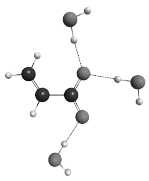
Glycolic acid (exptl. pK_a = 3.84)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			2.64	1.20
1			2.79	1.05
1			3.50	0.34
2			4.03	-0.19
2			3.30	0.54
2			1.94	1.90
2			2.67	1.17

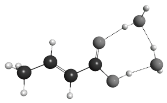
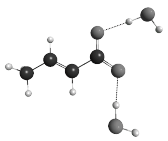
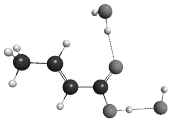
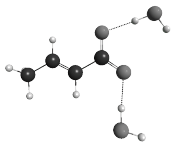
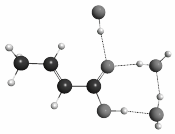
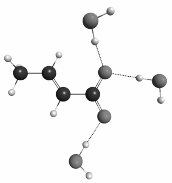
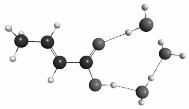
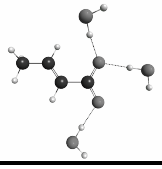
3			3.65	0.19
3			3.65	0.19
3			3.32	0.52

Acrylic acid (exptl. $pK_a = 4.26$)

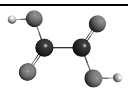
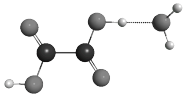
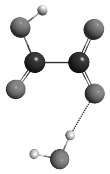
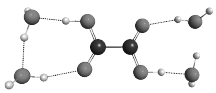
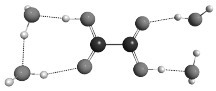
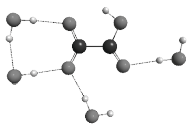
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			5.01	-0.75
1			5.47	-1.21
2			4.92	-0.66
2			5.31	-1.05
2			3.76	0.50
3			3.85	0.41

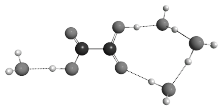
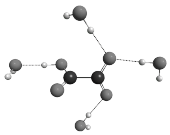
3			4.55	-0.29

Crotonic acid (exptl. $pK_a = 4.69$)

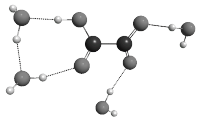
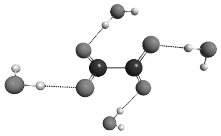
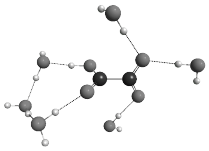
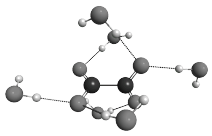
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.41	-1.72
2			6.16	-1.47
2			4.76	-0.07
3			4.75	-0.06
3			5.08	-0.39

Oxalic acid (exptl. $pK_a^1 = 1.23$)

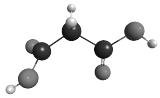
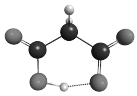
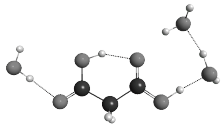
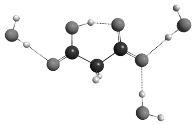
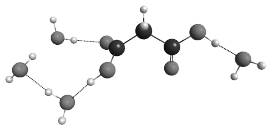
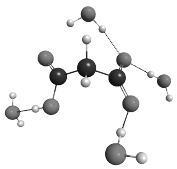
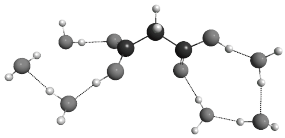
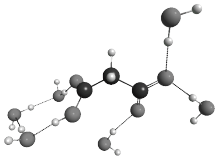
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			-3.08	4.31
1			-1.99	3.22
2 + 2			0.93	0.30
2 + 2 / 3 + 1			0.86	0.37

3 + 1			1.44	-0.21

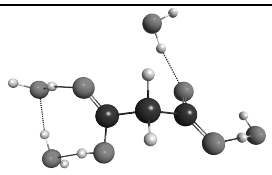
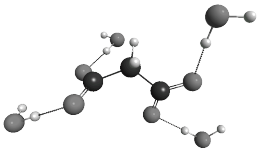
Oxalic acid (exptl. $pK_a^2 = 4.19$)

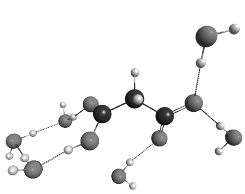
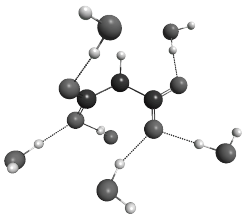
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
2 + 2			4.60	-0.41
3 + 3			4.62	-0.43

Malonic acid (exptl. $pK_a^1 = 2.83$)

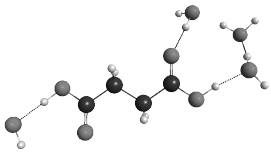
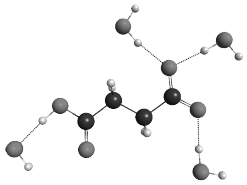
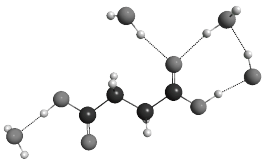
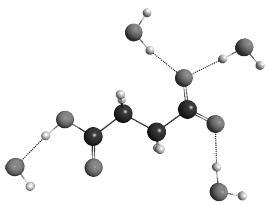
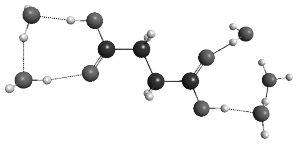
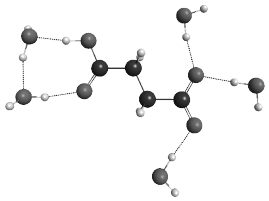
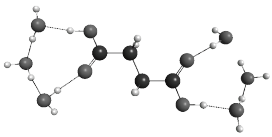
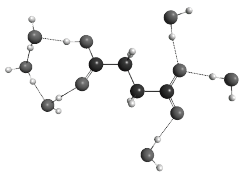
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			-3.84	6.67
2 + 1			0.18	2.65
3 + 1			3.39	-0.56
3 + 1			4.60	-1.77

Malonic Acid (exptl. $pK_a^2 = 5.69$)

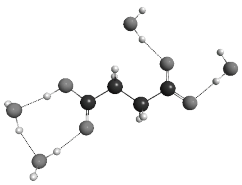
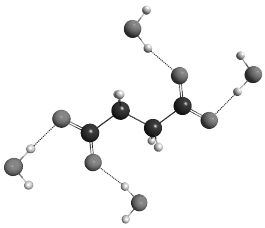
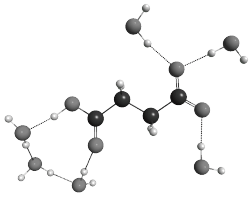
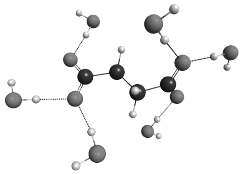
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
2 + 2			5.48	0.21

3 + 3			5.70	-0.01

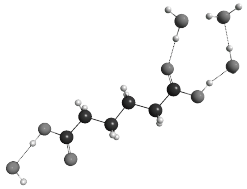
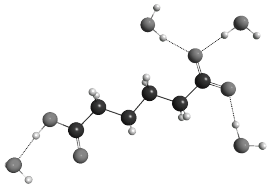
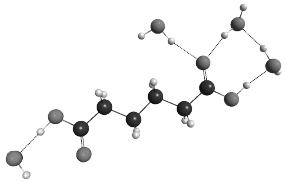
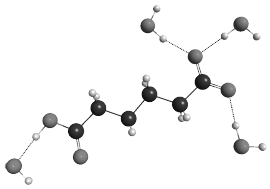
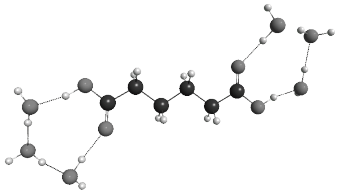
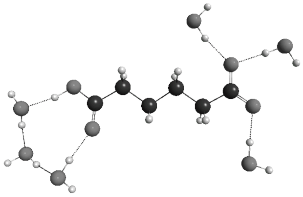
Succinic acid (exptl. pK_a^1 = 4.16)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 + 1			5.04	-0.88
3 + 1			4.00	0.16
3 + 2			4.94	-0.78
3 + 3			4.95	-0.79

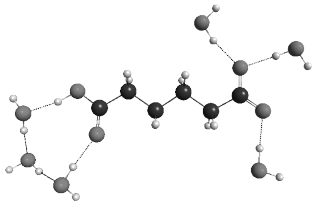
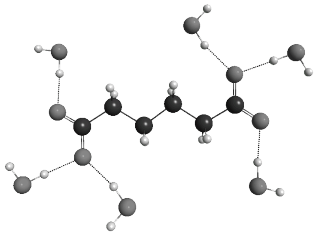
Succinic acid (exptl. pK_{a2} = 5.61)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
2 + 2			7.74	-2.13
3 + 3			5.51	0.10

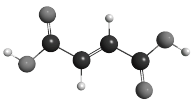
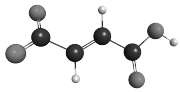
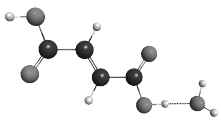
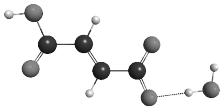
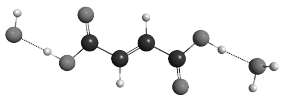
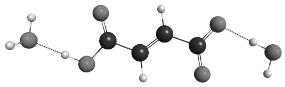
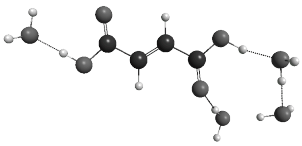
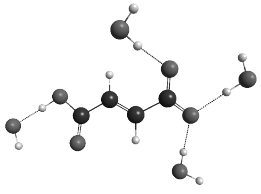
Adipic acid (exptl. pK_a^1 = 4.43)

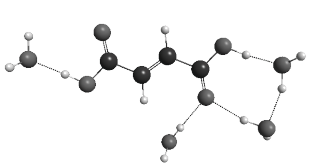
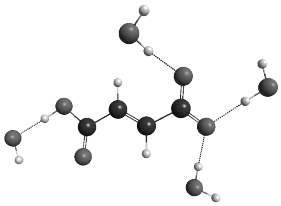
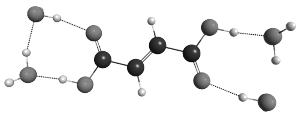
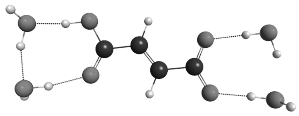
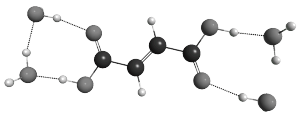
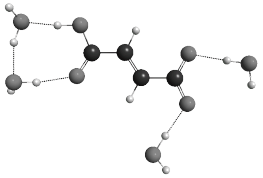
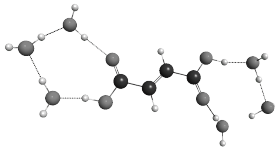
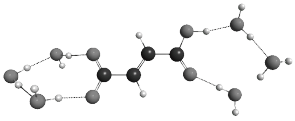
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 + 1			5.23	-0.80
3 + 1			4.76	-0.33
3 + 3			5.18	-0.75

Adipic acid (exptl pK_a^2 = 5.41)

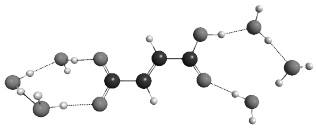
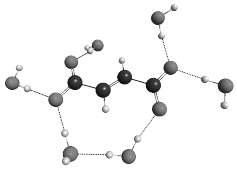
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 +3			5.85	-0.44

Fumaric acid (exptl. pK_a^1 = 3.03)

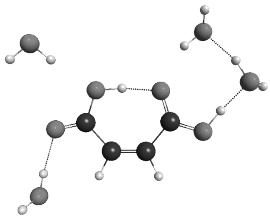
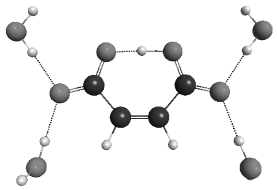
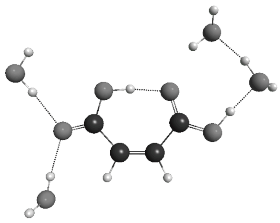
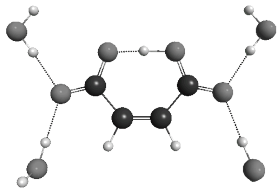
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			2.72	0.31
1			3.61	-0.58
1 + 1			3.93	-0.90
3 + 1			4.16	-1.13

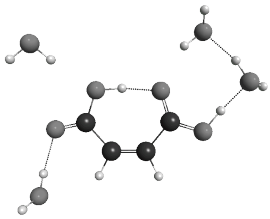
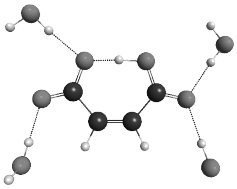
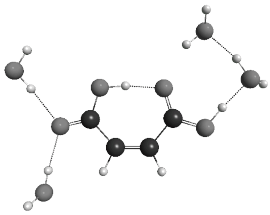
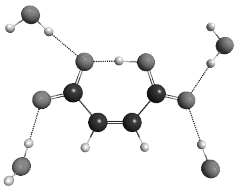
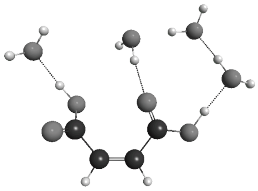
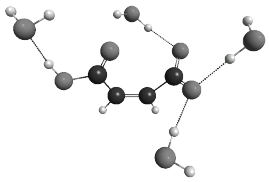
3 + 1			2.82	0.21
2 + 2			3.82	-0.79
2 + 2			4.20	-1.17
3 + 3			3.78	-0.75

Fumaric acid (exptl. $pK_a^2 = 4.44$)

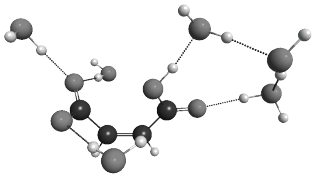
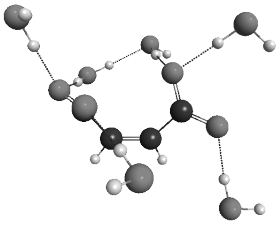
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 + 3			4.68	-0.24

Maleic acid (exptl. $pK_a^1 = 1.83$)

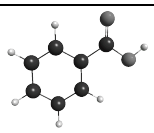
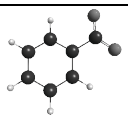
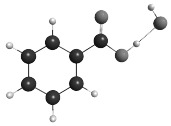
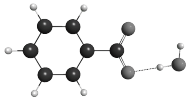
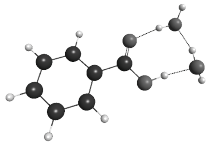
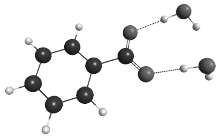
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
2 + 2			-2.29	4.12
2 + 2			-1.78	3.61

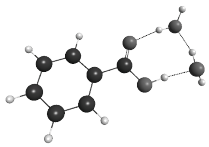
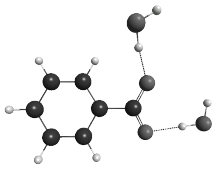
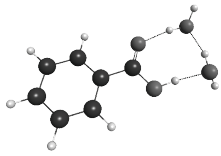
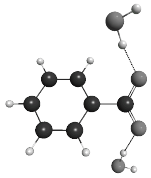
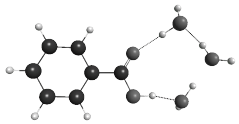
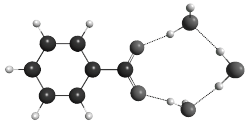
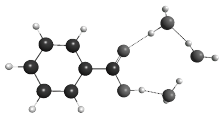
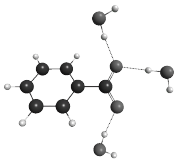
2 + 2			-1.48	3.31
2 + 2			-0.97	2.80
3 + 1			2.57	-0.74

Maleic acid (exptl. $pK_a^2 = 6.07$)

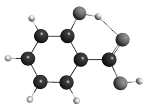
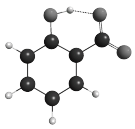
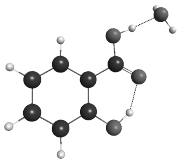
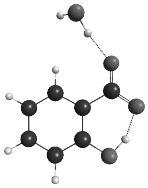
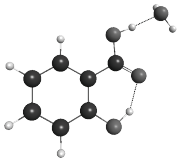
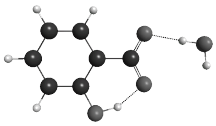
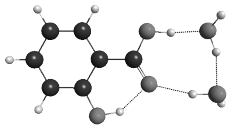
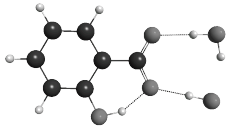
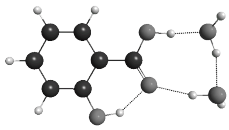
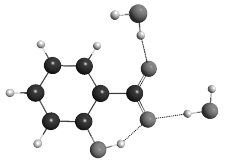
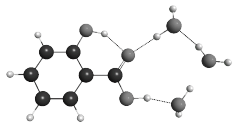
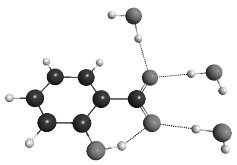
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 + 3			6.04	0.03

Benzoic acid (exptl. $pK_a = 4.2$)

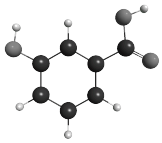
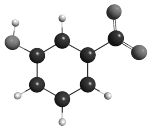
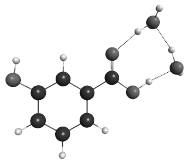
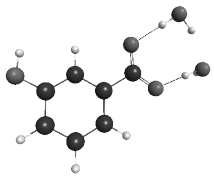
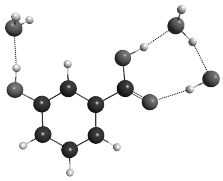
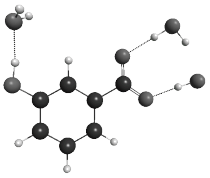
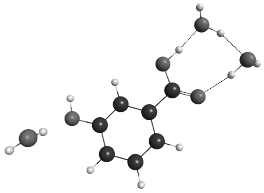
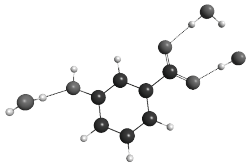
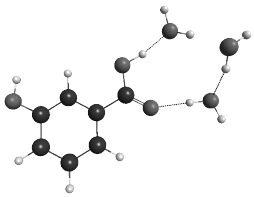
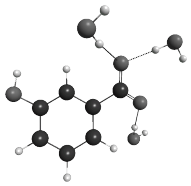
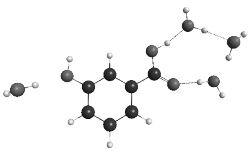
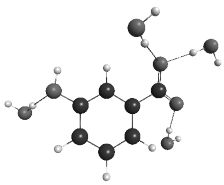
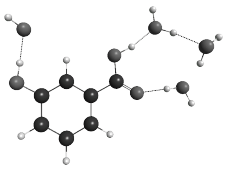
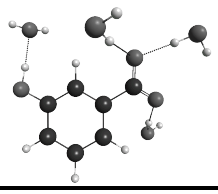
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			4.70	-0.50
1			5.18	-0.98
2			4.47	-0.27

2			5.59	-1.39
2			4.84	-0.64
3			4.88	-0.68
3			4.70	-0.50

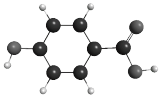
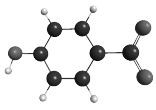
Salicylic acid (exptl. pK_a = 2.98)

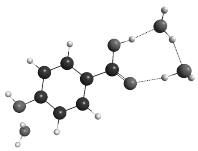
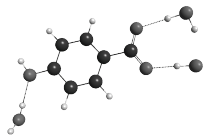
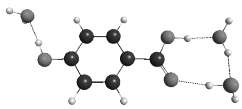
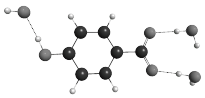
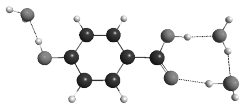
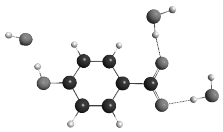
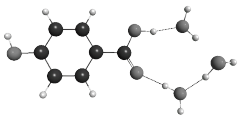
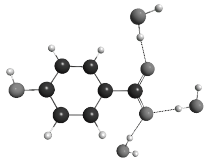
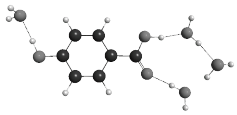
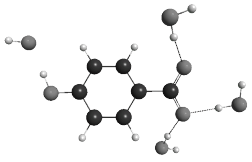
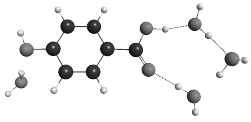
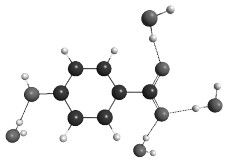
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			2.22	0.76
1			3.22	-0.24
1			3.01	-0.03
2			2.66	0.32
2			3.47	-0.49
3			2.30	-0.68

m-Hydroxybenzoic acid (exptl. pK_a = 4.08)

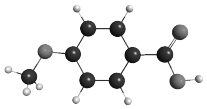
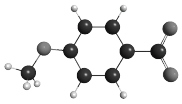
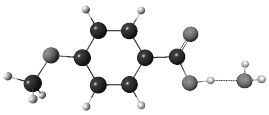
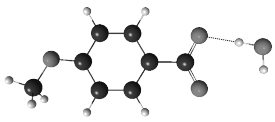
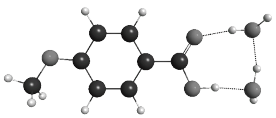
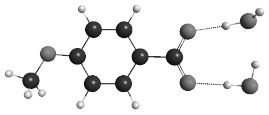
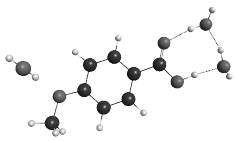
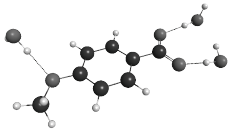
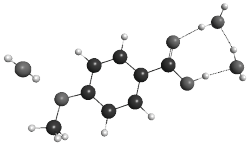
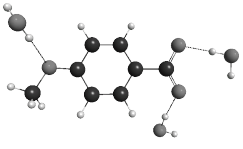
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			4.56	-0.48
2			4.60	-0.52
2 + 1			4.61	-0.53
2 + 1			3.87	0.21
3			4.52	-0.44
3 + 1			3.92	0.16
3 + 1			5.14	-1.06

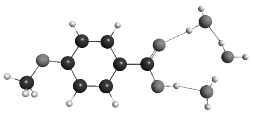
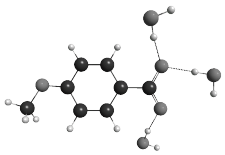
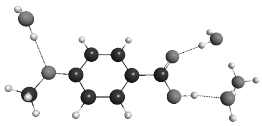
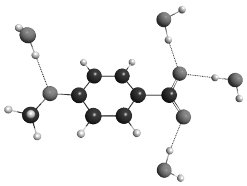
p-Hydroxybenzoic acid (exptl. pK_a = 4.58)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.25	-1.67

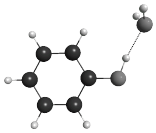
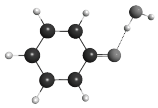
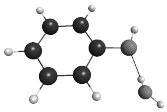
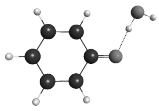
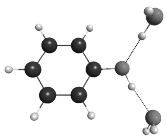
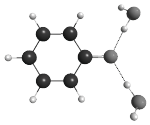
2 + 1			4.80	-0.22
2 + 1			5.76	-1.18
2 + 1			6.75	-2.17
3			5.04	-0.46
3 + 1			5.39	-0.81
3 + 1			4.45	0.13

p-Methoxybenzoic acid (exptl. $pK_a = 4.5$)

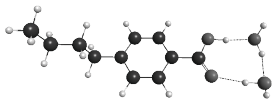
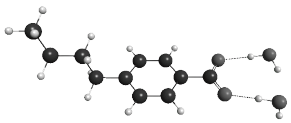
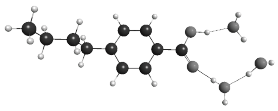
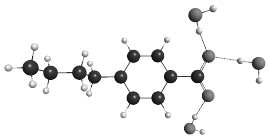
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.50	-2.00
1			6.31	-1.81
2			5.47	-0.97
2 + 1			4.72	-0.22
2 + 1			5.75	-1.20

3			5.37	-0.87
3 + 1			5.24	-0.74

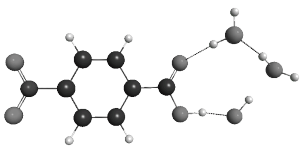
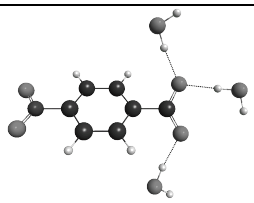
Phenol (exptl. pK_a = 9.82)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			11.22	1.40
1			10.38	0.56
1			8.91	0.91
2			8.07	1.75

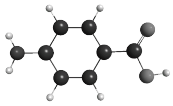
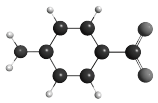
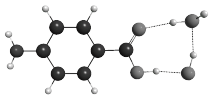
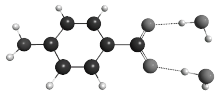
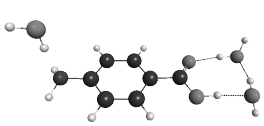
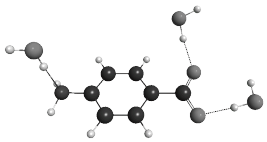
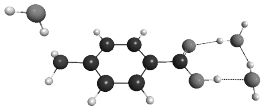
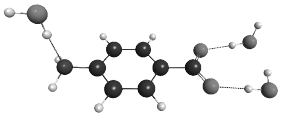
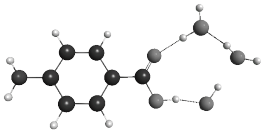
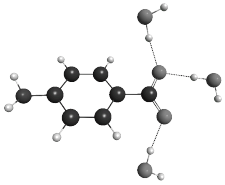
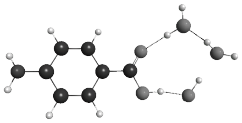
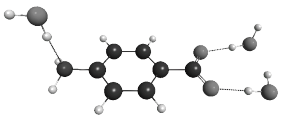
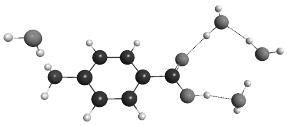
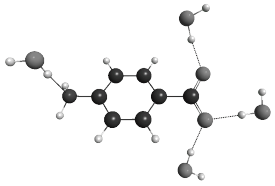
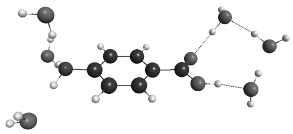
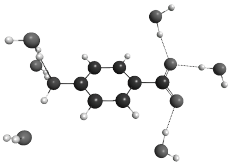
p-Butylbenzoic acid (exptl. pK_a = 4.47)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
2			5.16	-0.69
3			5.02	-0.55

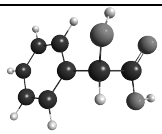
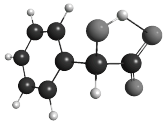
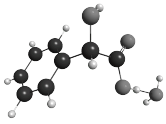
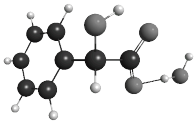
p-Nitrobenzoic acid (exptl. pK_a = 3.4)

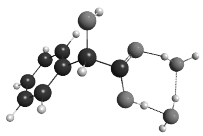
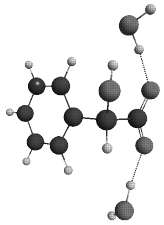
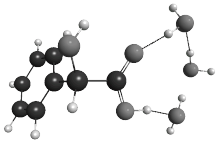
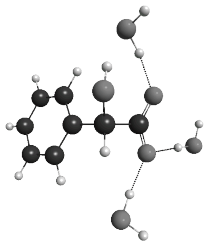
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3			3.19	0.21

p-Aminobenzoic acid (exptl. pK_a = 4.92)

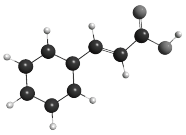
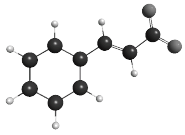
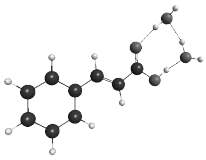
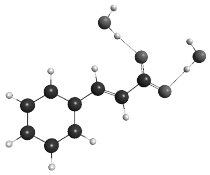
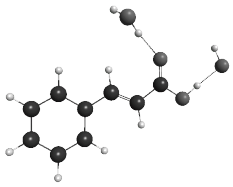
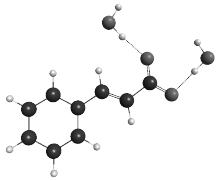
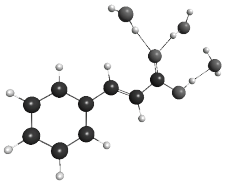
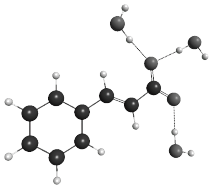
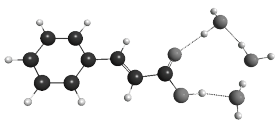
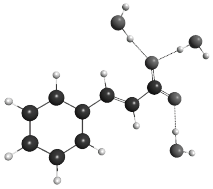
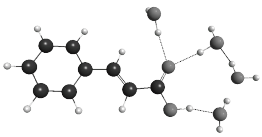
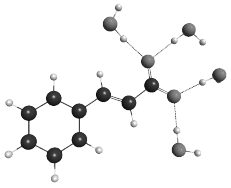
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			7.38	-2.46
2			6.39	-1.47
2 + 1			6.77	-1.85
2 + 1			5.74	-0.82
3			6.31	-1.39
3/ 2+1			6.00	-1.08
3 + 1			5.86	-0.94
3 + 3			5.90	-0.98

Mandelic acid (exptl. pK_a = 3.41)

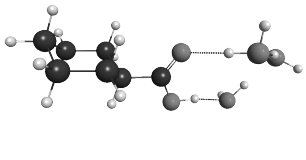
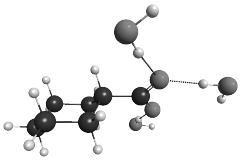
S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			2.34	1.07
1			3.10	0.31

2			3.68	-0.27
3			3.11	0.30

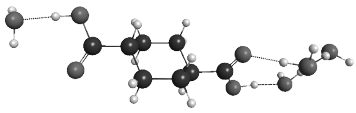
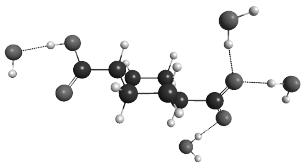
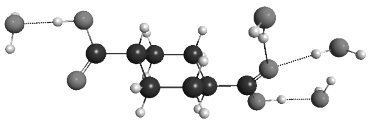
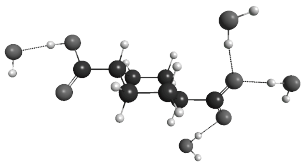
Cinnamic acid (exptl. pK_a = 4.44)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			6.22	-1.78
2			6.45	-2.01
2			5.02	-0.58
3			4.68	-0.24
3			5.40	-0.96
4			3.28	1.16

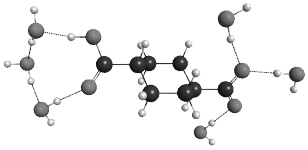
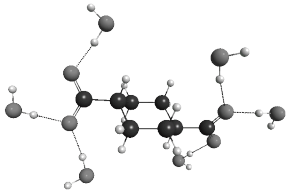
Cyclohexanecarboxylic acid (exptl. pK_a = 4.9)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3			5.62	0.72

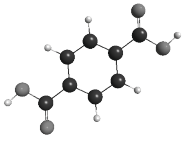
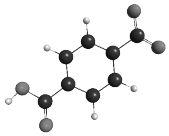
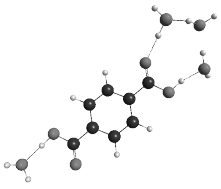
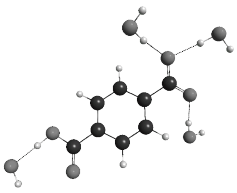
Cyclohexanedicarboxylic acid (exptl. pK_a^1 = 4.18)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			5.93	1.75
3 + 1			4.96	0.78
3 +1			4.46	0.28

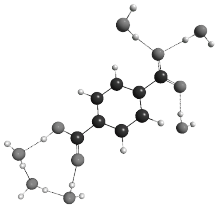
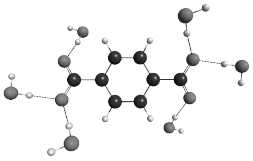
Cyclohexanedicarboxylic acid (exptl. pK_a^2 = 5.42)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 + 3			6.17	-0.75

Terephthalic Acid (exptl. pK_a^1 = 3.51)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
0			4.14	0.63
3 + 1			4.07	0.56

Terephthalic Acid (exptl. $pK_a^2 = 4.4$)

S_D	Acid cluster	Anion cluster	Calculated pK_a	ΔpK_a(exptl. – calc.)
3 + 3			5.19	-0.79

Appendix B

Changes identified to fix the dmulti OCE routine (in red), in the COSOCE subroutine in cosmo.src.

```
C   FOR THE DISTRIBUTED MULTIPOLAR OUTLYING CHARGE CORRECTION WE
C   NEED TO RE-EVALUATE THE REAL POTENTIAL ON THE CAVITY.
C   THE CORRECTION IS THE DIFFERENCE BETWEEN THE COSMO POTENTIAL
C   AND THE REAL POTENTIAL.
C   THIS IS DONE HERE ONLY FOR SCF - FOR MP2 THE OUTLYING CHARGE
C   CORRECTION IS SKIPPED AND DONE IN MP2GRD
C
      IF(OUTCHG.EQ.DMULTI) THEN
C
C     CALCULATE QVCORR - USING REAL POTENTIAL FROM GAMESS
C
       CALL COSPOT(QVPOT)
C
        QVCORR=0.0D+00
        DO 6 I=1,NQS
C FIX FOR UNIT FOR DMULTI
c         QVPOT(I)=QVPOT(I)*TOANGS
         QVCORR = QVCORR + QVPOT(I)*QSCNET(I)
  6     CONTINUE
C
C     CALCULATE "OUTLYING CHARGE" (EOC1): -QV + QV'
C
        EOC1=-QVCORR+QVCOSMO
      END IF
C
C   PEDIFF6 IS NOW THE 'CORRECTED' TOTAL ENERGY AND
C   WE PUT IT BACK INTO ENRTOADD
C
      PEDIFF6=ENRTOADD+EOC1
      ENRTOADD=PEDIFF6
C
C     RESET SOME ESSENTIAL PARAMETERS:
C
      ITRIP=0
      USEPS=.FALSE.
      ELAST=0.0D+00
      RETURN
C
 100    END
```

Missing call to obtain the potential QVPOT

Units are already correct. Comment out change to Angstrom

Appendix C

The following error was found in the SCFLIB.SRC affecting calculations of atoms.
Current GAMESS code:

```
      CDUM = ZERO
      IF(ISEPS  .AND.  MPCTYP.EQ.NONE  .AND.  ICORR.EQ.1) THEN
        IF (COSBUG) THEN
         WRITE(IW,*)'INSIDE SCFLIB, EXTRA NUCLEAR LOOP, N,NPS=',N,NPS
        ENDIF
        DO 122 I=1,NPS
          DO 124 J=1,NATOMS
            RR=ZERO
            DO 126 K=1,3
              RR=RR+(CORZAN(K,I)-C(K,J))**2
 126          CONTINUE
            CDUM = CDUM + QSCNET(I)*Z(J)/SQRT(RR)
 124        CONTINUE
 122      CONTINUE
        IF (COSBUG) THEN
         WRITE(IW,*)'NUCLEAR CHARGE (REPNUC), NO NUCLEAR-SAS:',REPNUC
         WRITE(IW,*)'NUCLEAR-SAS CONTRIBUTION (CDUM):',CDUM
        ENDIF
        REPNUC=REPNUC+CDUM
        IF (COSBUG) THEN
         WRITE(IW,*)'TOTAL VALUE OF NUCLEAR-CHARGE COMPONENT:',REPNUC
        ENDIF
      END IF
C
C          ADD NUCLEAR CONTRIBUTION FROM ELECTRIC FIELD
C
  310 CONTINUE
```

Corrected Code:

```
  310 CONTINUE
      CDUM = ZERO
      IF(ISEPS  .AND.  MPCTYP.EQ.NONE  .AND.  ICORR.EQ.1) THEN
        IF (COSBUG) THEN
         WRITE(IW,*)'INSIDE SCFLIB, EXTRA NUCLEAR LOOP, N,NPS=',N,NPS
        ENDIF
        DO 122 I=1,NPS
          DO 124 J=1,NATOMS
            RR=ZERO
            DO 126 K=1,3
              RR=RR+(CORZAN(K,I)-C(K,J))**2
 126          CONTINUE
            CDUM = CDUM + QSCNET(I)*Z(J)/SQRT(RR)
 124        CONTINUE
 122      CONTINUE
        IF (COSBUG) THEN
         WRITE(IW,*)'NUCLEAR CHARGE (REPNUC), NO NUCLEAR-SAS:',REPNUC
         WRITE(IW,*)'NUCLEAR-SAS CONTRIBUTION (CDUM):',CDUM
        ENDIF
        REPNUC=REPNUC+CDUM
        IF (COSBUG) THEN
         WRITE(IW,*)'TOTAL VALUE OF NUCLEAR-CHARGE COMPONENT:',REPNUC
        ENDIF
      END IF
```

Appendix D

Changes to add new radii options
In COSMIN:
1) Add new keywords to the COSDAT COMMON block – must be added to all COSDAT blocks throughout all source files

```
      LOGICAL COSBUG,COSWRT,DCOSMO,PRFCND,IOUTCH,ISOCAV
      COMMON /COSDAT/ SE2,SECORR,QVCOSMO,ELAST,EMP2COS,EMP2LAST,
     *                COSVOL,COSSAR,EDIEL,EOC1,DEOC_RS,SUMQSC,
     *                SUMQSCOLD,ZSUM,ZSUM2,ZSUM3,FEPSI,RDS,DISEX2,
     *                EPSI,COSRAD,VDWFAC,COSDEN,DISEX,OUTCHG,
     *                VDWRAD,EDIEL_SAVE,
     *                MAXNPS,ICORR,ITRIP,NQS,MP2TRIP,MP2ITER,
     *                ICFREQ,NSPA,NSPH,NPSD,NPS,NPS2,NDEN,NPSPHER,
     *                COSBUG,COSWRT,DCOSMO,PRFCND,IOUTCH,ISOCAV
```

2) specify new data groups

```
      DATA VDWKLM,VDWBON,VDWALV,VDWEMS/8HVDWKLM  ,8HVDWBON  ,
     *                                 8HVDWALV  ,8HVDWEMS  /
```

3) Add to QNAM and specify type of data input

```
      PARAMETER (NNAM=13)
      DIMENSION QNAM(NNAM),KQNAM(NNAM)
C
      DATA DMULTI,DBLCAV,OCENON/8HDMULTI  ,8HDBLCAV  ,8HOCENON  /
      DATA VDWKLM,VDWBON,VDWALV,VDWEMS/8HVDWKLM  ,8HVDWBON  ,
     *                                 8HVDWALV  ,8HVDWEMS  /
      DATA BLANK/8H        /
      DATA COSGMS/8HCOSGMS  /
      DATA QNAM/8HEPSI    ,8HNSPA    ,8HCOSRAD  ,
     *          8HDISEX   ,8HOUTCHG  ,8HVDWRAD  ,
     *          8HCOSBUG  ,8HCOSWRT  ,8HDCOSMO  ,
     *          8HPRFCND  ,8HISOCAV  ,8HCOSDEN  ,
     *          8HVDWFAC  /
      DATA KQNAM /3,1,3,3,5,5,0,0,0,0,0,3,3/
```

4) Set defaults

```
C INITIALIZATION OF SOME PARAMETERS
C
      ITRIP = 0
      MP2TRIP = 0
      ICFREQ = 0
      EPSI = 0.0D+00
      COSRAD = 1.2D+00
      NSPA  = 92
      DISEX = 10.0D+00
      OUTCHG = BLANK
      VDWRAD = BLANK
      COSBUG = .FALSE.
      COSWRT = .FALSE.
      DCOSMO = .FALSE.
      PRFCND = .FALSE.
      ISOCAV = .FALSE.
      COSDEN = 0.001D+00
      VDWFAC = 1.2D+00
```

5) Include in call to NAMEIO to read from $COSGMS and set options

```
C READ $COSGMS.
C
      JRET = 0
      CALL NAMEIO(IR,JRET,COSGMS,NNAM,QNAM,
     *           KQNAM,EPSI,NSPA,COSRAD,DISEX,
     *           OUTCHG,VDWRAD,COSBUG,COSWRT,DCOSMO,
     *           PRFCND,ISOCAV,COSDEN,VDWFAC,0,
     *           0,0,0,0,0,   0,0,0,0,0,
     *           0,0,0,0,0,   0,0,0,0,0,
     *           0,0,0,0,0,   0,0,0,0,0,
     *           0,0,0,0,0,   0,0,0,0,0,
     *           0,0,0,0,0,   0,0,0,0,0)
      IF(JRET.EQ.2) THEN
        IF (MASWRK) WRITE(IW,9028)
        CALL ABRT
      END IF
      IF(JRET.EQ.1) RETURN
      NERR = 0
C
C
C     SET OPTIONS AND CHECK SPELLING FOR $COSMO
C
      OK = .FALSE.
      IF(OUTCHG.EQ.BLANK) OUTCHG = DBLCAV
      IF(OUTCHG.EQ.OCENON)   OK=.TRUE.
      IF(OUTCHG.EQ.DMULTI)   OK=.TRUE.
      IF(OUTCHG.EQ.DBLCAV)   OK=.TRUE.
      IF(.NOT.OK) THEN
         IF (MASWRK) WRITE(IW,9010) 'OUTCHG',OUTCHG
         NERR = NERR+1
      END IF
C
C ba set options for VDWRAD
      OK = .FALSE.
      IF(VDWRAD.EQ.BLANK) VDWRAD = VDWBON
      IF(VDWRAD.EQ.VDWKLM) OK=.TRUE.
      IF(VDWRAD.EQ.VDWALV) OK=.TRUE.
      IF(VDWRAD.EQ.VDWEMS) OK=.TRUE.
      IF(VDWRAD.EQ.VDWBON) OK=.TRUE.
      IF(.NOT.OK) THEN
         IF(MASWRK) WRITE(IW,9010) 'VDWRAD',VDWRAD
         NERR = NERR+1
      ENDIF
```

6) Add to print statements

```
C  PRINT $COSGMS OPTIONS
C
       IF(PRFCND) THEN
         IF(MASWRK) WRITE(IW,9057) NSPA,COSRAD,DISEX,VDWFAC,
     *                             COSDEN,OUTCHG,VDWRAD,COSBUG,
     *                             COSWRT,DCOSMO,ISOCAV
         IF(EPSI.NE.0.0D+00.AND.MASWRK) WRITE(IW,9090)
       ELSE
         IF(MASWRK) WRITE(IW,9058) EPSI,NSPA,COSRAD,DISEX,VDWFAC,
     *                             COSDEN,OUTCHG,VDWRAD,COSBUG,
     *                             COSWRT,DCOSMO,ISOCAV
       END IF
  9057 FORMAT(/5X,'$COSGMS OPTIONS'/5X,15(1H-)/
     *  5X,7HEPSI  =,'   INFTY',5X,7HNSPA  =,I8,5X,7HCOSRAD=,F8.2/
     *  5X,7HDISEX =,F8.2,5X,7HVDWFAC=,F8.2,5X,7HCOSDEN=,F7.5/
     *  5X,7HOUTCHG=,A8,5X,7HVDWRAD=,A8,5X,7HCOSBUG=,L8/
     *  5X,7HCOSWRT=,L8,5X,7HDCOSMO=,L8,5X,7HISOCAV=,L8)
  9058 FORMAT(/5X,'$COSGMS OPTIONS'/5X,15(1H-)/
     *  5X,7HEPSI  =,F8.3,5X,7HNSPA  =,I8,5X,7HCOSRAD=,F8.2/
     *  5X,7HDISEX =,F8.2,5X,7HVDWFAC=,F8.2,5X,7HCOSDEN=,F7.5/
     *  5X,7HOUTCHG=,A8,5X,7HVDWRAD=,A8,5X,7HCOSBUG=,L8/
     *  5X,7HCOSWRT=,L8,5X,7HDCOSMO=,L8,5X,7HISOCAV=,L8)
```

In SVINIT:

1) Dimension new data blocks

```
      DIMENSION RKLAMT(53), REMSL(53), RBOND(53),
     * RALVZ(53), USEVDW(53)
      DIMENSION CCORE(107)
```

2) Add radii data info

```
   DATA VDWKLM,VDWBON,VDWALV,VDWEMS/8HVDWKLM  ,8HVDWBON  ,
  *                                 8HVDWALV  ,8HVDWEMS  /
```

3) Define new radii blocks

```
      DATA RKLAMT   /1.30D+00, 1.0D+00, 1.80D+00, 999.0D+00, 2.08D+00,
     1 2.00D+00,   1.83D+00,  1.72D+00,  1.72D+00,  1.50D+00,
     2 2.31D+00,   999.0D+00, 2.05D+00,  2.10D+00,  1.90D+00,
     3 2.16D+00,   2.05D+00,  999.0D+00, 2.31D+00,  1.74D+00,
     4 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     5 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 1.39D+00,
     6 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 2.160D+00,
     7 999.0D+00,  2.44D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     8 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     9 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     * 999.0D+00,  999.0D+00, 2.32D+00 /
C
C The following is the vdw radii taken from Emsley 1989
      DATA REMSL   /1.20D+00, 1.22D+00, 999.0D+00, 999.0D+00, 2.08D+00,
     1 1.85D+00, 1.54D+00, 1.40D+00, 1.35D+00, 1.60D+00,
     2 2.31D+00, 999.0D+00, 2.05D+00, 2.00D+00, 1.90D+00,
     3 1.85D+00, 1.81D+00, 1.91D+00, 2.31D+00, 999.0D+00,
     4 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     5 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     6 999.0D+00, 999.0D+00, 2.00D+00, 2.00D+00, 1.95D+00,
     7 1.98D+00, 2.44D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     8 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     9 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     * 999.0D+00,  999.0D+00, 2.15D+00 /
C
C Bondi radii or otherwise specified from pcm code
      DATA RBOND   /1.20d+00, 1.40d+00, 1.82d+00, 1.45d+00, 1.80d+00,
     1 1.70d+00, 1.55d+00, 1.52d+00, 1.47d+00, 1.54d+00,
     2 2.27d+00, 1.73d+00, 2.30d+00, 2.10d+00, 1.80d+00,
     3 1.80d+00, 1.75d+00, 1.88d+00, 2.75d+00, 999.0d+00,
     4 999.0D+00,  999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     5 999.0d+00, 999.0d+00, 1.63d+00, 1.40d+00, 1.39d+00,
     6 1.87d+00, 2.19d+00, 1.85d+00, 1.90d+00, 1.85d+00,
     7 2.02d+00, 999.0d+00, 999.0d+00, 999.0d+00, 999.0d+00,
     8 999.0d+00, 999.0d+00, 999.0d+00, 999.0d+00, 999.0d+00,
     9 1.63d+00, 1.72d+00, 1.58d+00, 1.93d+00, 2.17d+00,
     * 999.0d+00, 2.06d+00, 1.98d+00 /
c 2.16d+00, 999.0d+00,
c     * 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
c     * 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
c     * 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
c     * 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
c     * 999.0D+00, 999.0D+00, 1.75d+00, 1.66d+00, 1.55d+00,
c     * 1.96d+00, 2.02d+00, 999.0D+00, 999.0D+00, 999.0D+00,
c     * 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
c     * 999.0D+00, 1.86d+00 /
c
C Alvarez 2013 radii
      DATA RALVZ   /1.20d+00, 1.4d+00, 1.81d+00, 999.0d+00, 999.0d+00,
     1 1.70d+00, 1.55d+00, 1.52d+00, 1.47d+00, 1.54d+00,
     2 2.27d+00, 1.73d+00, 999.0d+00, 2.22d+00, 1.80d+00,
     3 1.80d+00, 1.75d+00, 1.76d+00, 2.75d+00, 999.0d+00,
     4 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     5 999.0D+00, 999.0D+00, 1.63d+00, 1.40d+00, 1.39d+00,
     6 1.87d+00, 999.0D+00, 1.85d+00, 1.90d+00, 1.83d+00,
     7 2.02d+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     8 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00, 999.0D+00,
     9 1.63d+00, 1.72d+00, 1.62d+00, 1.93d+00, 2.17d+00,
     * 999.0d+00, 2.00d+00, 1.98d+00/
```

4) Read in radii for atoms 1 - 53.

```
      NPPAX = NPPA
      CALL AOLIM()
C
      IF(VDWRAD.EQ.VDWKLM) THEN
         DO 10 I=1,53
             USEVDW(I)=RKLAMT(I)
   10 CONTINUE
      ELSEIF(VDWRAD.EQ.VDWBON) THEN
         DO 11 I=1,53
             USEVDW(I)=VDWFAC*RBOND(I)
   11 CONTINUE
      ELSEIF(VDWRAD.EQ.VDWEMS) THEN
         DO 12 I=1,53
             USEVDW(I)=VDWFAC*REMSL(I)
   12 CONTINUE
      ELSEIF(VDWRAD.EQ.VDWALV) THEN
         DO 13 I=1,53
             USEVDW(I)=VDWFAC*RALVZ(I)
   13 CONTINUE
      ENDIF
C
C
C  CORRESPING RADII TO ATOMIC NUMBER
C
```

Bibliography

[1] R. Shivapurkar, D. Jeannerat, *Anal. Methods* **2011**, *3*, 1316.
[2] P. Gilli, L. Pretto, V. Bertolasi, G. Gilli, *Acc. Chem. Res.* **2009**, *42*, 33-44.
[3] L. Gregerson, K. K. Baldridge, *Helvetica Chimica Acta* **2003**, *86*, 4112.
[4] M. Schmidt, K. K. Baldridge, J. A. Boatz, S. Elbert, M. Gordon, J. H. Jensen, S. Koeski, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comp. Chem.* **1993**, *14*, 1347-1363.
[5] C. Reichardt, *Solvents and Solvent Effects in Organic Chemistry*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, **2003**.
[6] H. Ohtaki, *Monatshefte für Chemie* **2001**, *132*, 1237-1268.
[7] K. J. Laidler, *Pure & Appl. Chem.* **1990**, *62*, 2221-2226.
[8] H. Ohtaki, *Coord. Chem. Rev.* **1999**, *185-186*, 735-759.
[9] R. M. Stratt, M. Maroncelli, *J. Phys. Chem.* **1996**, *100*, 12981-12996.
[10] I. Topol, G. Tawa, S. Burt, A. Rashin, *J. Chem . Phys.* **1999**, *111*, 10998.
[11] S. R. Pruitt, M. A. Addicoat, M. A. Collins, M. S. Gordon, *Phys. Chem. Chem. Phys.* **2012**, *14*, 7752.
[12] A. Klamt, *COSMO-RS From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, Elsevier, The Netherlands, **2005**.
[13] a C. G. Zhan, J. Bentley, D. M. Chipman, *J. Chem. Phys.* **1998**, *108*, 177; b J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999-3094.
[14] C. Amovilli, V. Barone, R. Cammi, E. Cancès, M. Cossi, B. Mennucci, C. S. Pomelli, J. Tomasi, *Advances in Quantum Chemistry* **1999**, *32*, 227-261.
[15] C. J. Cramer, D. G. Truhlar, *Acc. Chem. Res.* **2008**, *41*, 760-768.
[16] M. J. Vilkas, C.-G. Zhan, *J. Chem. Phys.* **2008**, *129*, 194109.
[17] A. Klamt, G. Schüürmann, *J. Chem. Soc. Perkin Trans. 2* **1993**, 799-805.
[18] A. Klamt, V. Jonas, *J. Chem. Phys.* **1996**, *105*, 9972.
[19] E. Cances, B. Mennucci, *J. Chem. Phys.* **2001**, *115*, 6130-6135.
[20] C. C. Pye, T. Ziegler, *Theor. Chem. Acc.* **1999**, *101*, 396-408.
[21] M. Cossi, N. Rega, G. Scalmani, V. Barone, *J. Chem. Phys.* **2001**, *114*, 5691-5701.
[22] a V. Barone, M. Cossi, J. Tomasi, *J. Chem. Phys.* **1997**, *107*, 3210; b A. Pomogaeva, D. M. Chipman, *J. Phys. Chem. A* **2013**, *117*, 5812-5820.
[23] J. B. Foresman, T. A. Keith, K. B. Wiberg, J. Snoonian, M. J. Frisch, *J. Phys. Chem.* **1996**, *100*, 16098.
[24] a K. B. Wiberg, T. A. Keith, M. J. Frisch, M. Murcko, *J. Phys. Chem.* **1995**, *99*, 9072; b K. B. Wiberg, Rablen, P. R., Rush, D. J., Keith, T.A. , *J. Am. Chem. Soc.* **1995**, *117*, 4261.
[25] J. Bentley, *J. Phys. Chem. A* **1998**, *102*, 6043-6051.
[26] C.-G. Zhan, D. M. Chipman, *J. Chem. Phys.* **1998**, *109*, 10543.
[27] C. J. Cramer, D. G. Truhlar, *J. Chem. Rev.* **1999**, *99*, 2161-2200.
[28] J. Liu, C. P. Kelly, A. C. Goren, A. V. Marenich, C. J. Cramer, D. G. Truhlar, C.-G. Zhan, *J. Chem. Theory. Comput.* **2010**, *6*, 1109-1117.
[29] J. Pliego, J. Riveros, *J. Phys. Chem. A* **2001**, *105*, 7241-7247.
[30] a H. S. Frank, M. W. Evans, *J. Chem . Phys.* **1945**, *13*, 507; b H. S. Frank, W.-Y. Wen, *Discuss. Faraday Soc.* **1957**, *24*, 133.
[31] P. Bandyopadhyay, M. Gordon, *J. Chem . Phys.* **2000**, *113*, 1104.
[32] A. H. De Vries, P. T. Van Duijnen, A. H. Juffer, J. A. C. Rullmann, J. P. Dijkman, H. Merenga, B. T. Thole, *J. Comput. Chem.* **1995**, *16*, 37-55.

[33] K. K. Baldridge, A. Klamt, *J. Chem. Phys.* **1997**, *106*, 6622-6633.
[34] A. Klamt, V. Jonas, *J. Chem . Phys.* **1996**, *105*, 9972.
[35] R. Abramson, K. K. Baldridge, *Mol. Phys.* **2012**, *110*, 2401-2412.
[36] S. Grimme, *J. Comput. Chem.* **2006**, *27*, 1787-1799.
[37] a R. Peverati, K. K. Baldridge, *J. Chem. Theory Comput.* **2008**, *4*, 2030 - 2048; b R. Peverati, K. K. Baldridge, *J. Chem. Theory Comput.* **2009**, *5*, 2772-2786.
[38] R. Ditchfield, W. J. Hehre, J. A. Pople, *J. Chem. Phys.* **1971**, *54*, 724-728.
[39] a T. H. Dunning, *J. Chem. Phys.* **1970**, *53*, 2823; bT. H. Dunning, *J. Chem. Phys.* **1971**, *55*, 716.
[40] D. Rappoport, F. Furche, *J. Chem. Phys.* **2010**, *133*, 134105.
[41] J. Ho, M. Coote, *J. Chem. Theory Comput.* **2009**, 295-306.
[42] a L. A. Curtiss, K. Raghavachari, J. Pople, *J. Chem. Phys.* **1993**, *98*, 1293; b L. A. Curtiss, K. Raghavachari, P. Redfern, V. A. Rassolov, W. J. Pope, *J. Chem. Phys.* **1998**, *109*, 7764.
[43] G. A. Petersson, A. Bennett, T. G. Tensfeldt, M. A. Al-Laham, W. A. Shirley, J. Mantzaris, *J. Chem. Phys.* **1988**, *89*, 2193.
[44] a C. Kelly, C. Cramer, D. Truhlar, *J. Phys. Chem. B* **2006**, *110*, 16066; b M. Liptak, G. Shields, *Int. J. Quantum Chem.* **2001**, *85*, 727-741; c J. R. Pliego, *Chemical Physics Letters* **2003**, *367*, 145-149.
[45] a F. Eckert, M. Diedenhofen, A. Klamt, *Mol. Phys.* **2010**, *108*, 229-241; b J. Ho, A. Klamt, M. L. Coote, *J. Phys. Chem. A* **2010**, *114*, 13442-13444; c J. R. Pliego, J. M. Riveros, *J. Phys. Chem. A* **2002**, *106*, 7434-7439; d C. P. Kelly, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. A* **2006**, *110*, 2493-2499.
[46] A. Klamt, F. Eckert, M. Diedenhofen, M. Beck, *J. Phys. Chem. A* **2003**, *107*, 9380-9386.
[47] F. Ding, J. Smith, H. Wang, *Journal of Organic Chemistry* **2009**, *74*, 2679-2914.
[48] a J. Klicic, R. Friesner, S. Liu, W. Guida, *J. Phys. Chem. A* **2002**, *106*, 1327-1335; bK. R. Adam, *J. Phys. Chem. A* **2002**, *106*, 11963-11972; c S. Zhang, J. Baker, P. Pulay, *J. Phys. Chem. A* **2010**, *114*, 425-431; d M. Zimmermann, J. Tossell, *J. Phys. Chem. A* **2009**, *113*, 5105-5111; e Z. Jia, D. Du, Z. Zhou, A. Zhang, R. Hou, *Chemical Physics Letters* **2007**, *439*, 374-380.
[49] D. Chipman, *J. Phys. Chem. A* **2002**, *106*, 7413-7422.
[50] a J. Ho, M. Coote, *Theor. Chem. Acc.* **2010**, *125*, 3-21; b A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Chem. Theory Comput.* **2010**, *6*, 2829-2844.
[51] G. Klebe, *Vol. 2* (Eds.: H. B. Bürgi, J. D. Dunitz), Wiley-VCH, Weinheim, **1994**.
[52] a L. J. Henderson, *Am. J. Physiol.* **1906**, *21*, 173; b K. A. Hasselbalch, *Biochem. Z.* **1917**, *78*, 112.
[53] A. D. Becke, *J. Chem. Phys.* **1997**, *107*, 8554.
[54] C. Møller, M. S. Plesset, *Phys. Rev.* **1934**, *46*, 618-622.
[55] a A. D. McLean, G. S. Chandler, *J. Chem. Phys.* **1980**, *72*, 5639; b K. Raghavachari, J. S. Binkley, R. Seeger, J. A. Pople, *J. Chem. Phys.* **1980**, *72*, 650.
[56] K. K. Baldridge, V. Jonas, *J. Chem. Phys.* **2000**, *113*, 7511.
[57] A. Bondi, *J. Phys. Chem.* **1964**, *68*, 441.
[58] A. Klamt, V. Jonas, T. Bürger, C. W. Lohrenz, *J. Phys. Chem.* **1998**, *102*, 5074-5085.
[59] B. M. Bode, M. S. Gordon, *Mol. Graph. Mod.* **1999**, *16*, 133-138.

[60] J. Ho, M. L. Coote, *WIREs Comput. Molec. Sci.* **2011**, *1*, 649-660.
[61] a S. Zhang, *J. Comput. Chem.* **2011**, *33*, 517-526; b R. Casasnovas, D. Fernåndez, J. Ortega-Castro, J. Frau, J. Donoso, F. Muñoz, *Theor. Chem. Acc.* **2011**, *130*, 1-13; c M. Rupp, R. Körner, I. V. Tetko, *Com. Chem. High T. Scr.* **2011**, *14*, 307-327.
[62] a H. Lu, X. Chen, C. Zhan, *J. Phys. Chem. B* **2007**, *111*, 10599-10605; b B. Kallies, R. Mitzner, *J. Phys. Chem. B* **1997**, *101*, 2959-2967.
[63] a M. D. Liptak, K. C. Gross, P. G. Seybold, S. Feldgus, G. C. Shields, *J. Am. Chem. Soc.* **2002**, *124*, 6421-6427; bM. Liptak, G. Shields, *J. Am. Chem. Soc.* **2001**, *123*, 7314-7319; c M. Schmidt Am Busch, E.-W. Knapp, *Chem. Phys. Chem.* **2004**, *5*, 1513-1522; d V. Barone, R. Improta, N. Rega, *Theor. Chem. Acc.* **2004**, *111*, 237-245.
[64] a Y. Fu, L. Liu, R.-Q. Li, R. Liu, Q.-X. Guo, *J. Am. Chem. Soc.* **2004**, *126*, 814-822; b V. Bryantsev, M. Diallo, W. Goddard, *J. Phys. Chem. B* **2008**, *112*, 9709-9719; c J. R. Pliego, J. M. Riveros, *Phsy. Chem. Chem. Phys.* **2002**, *4*, 1622-1627; d X.-X. Wang, H. Fu, D.-M. Du, Z.-Y. Zhou, A.-G. Zhang, C.-F. Su, K.-S. Ma, *Chemical Physics Letters* **2008**, *460*, 339-342; e R. Casasnovas, J. Frau, J. Ortega-Castro, A. Salvà, *J. Mol. Struc. (Theochem)* **2009**, *912*, 5-12.
[65] C. Silva, E. da Silva, M. Nascimento, *J. Phys. Chem. A* **2000**, *104*, 2402-2409.
[66] D. M. Camaioni, C. A. Schwerdtfeger, *J. Phys. Chem. A* **2005**, *109*, 10795-10797.
[67] M. Tissandier, K. Cowen, W. Feng, E. Gundlach, M. Cohen, A. Earhart, J. Coe, T. Tuttle Jr, *J. Phys. Chem. A* **1998**, *102*, 7787-7794.
[68] A. Toth, M. Liptak, D. Phillips, G. Shields, *J. Chem. Phys.* **2001**, *114*, 4595-4606.
[69] G. Ramirez-Galicia, G. Perez-Caballero, M. Rubio, *J. Molecular Structure (Theochem)* **2001**, *542*, 1-6.
[70] D. M. Chipman, *J. Chem. Phys.* **2003**, *118*, 9937.
[71] a D. Du, M. Qin, Z. Zhou, A. Fu, *Int. J. Quantum Chem.* **2011**, *112*, 351-358; b A. V. Marenich, W. Ding, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. Lett.* **2012**, *3*, 1437-1442; c C. C. R. Sutton, G. V. Franks, G. da Silva, *J. Phys. Chem. B* **2012**, *116*, 11999-12006; d Y. C. Zheng, X. Chen, D. Zhao, H. Li, Y. Zhang, X. Xiao, *Fluid Phase Equilib.* **2012**, *313*, 148-155.
[72] a Y. Zhao, D. Truhlar, *Acc. Chem. Res.* **2008**, *41*, 157-167; b A. V. Marenich, C. Cramer, D. Truhlar, *J. Chem. Theory Comput.* **2008**, *4*, 877-887.
[73] R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople, *J. Chem. Phys.* **1980**, *72*, 650.
[74] C. Reichardt, *Solvents and Solvent Effects in Organic Chemistry*, VCH, Weinheim, **1988**.
[75] a T. Lee, M. McKee, *Phys. Chem. Chem. Phys.* **2011**, *13*, 10258-10269; b M. Śmiechowski, *J. Mol. Struct. (Theochem)* **2009**, *924-926*, 170-174.
[76] I. M. Alecu, J. Zheng, Y. Zhao, D. G. Truhlar, *J. Chem. Theory Comput.* **2010**, *6*, 2872-2887.
[77] M. L. Laury, M. J. Carlson, A. K. Wilson, *J. Comput. Chem.* **2012**, *33*, 2380-2387.
[78] M. L. Laury, S. E. Boesch, I. Haken, P. Sinha, R. A. Wheeler, A. K. Wilson, *J. Comput. Chem.* **2011**, *32*, 2339-2347.
[79] R. F. Ribeiro, A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* **2011**, *115*, 14556-14562.

[80] a L. M. Pratt, D. G. Truhlar, C. J. Cramer, S. R. Kass, J. D. Thompson, J. D. Xidos, *J. Org. Chem.* **2007**, *72*, 2962-2966; b Y. Zhao, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2008**, *10*, 2813.
[81] M. Cossi, B. Mennucci, J. Tomasi, *Chemical Physics Letters* **1994**, *228*, 165-170.
[82] a M. J, R. L. Jernigan, *Biopolymers* **1991**, *31*, 1615-1629; b V. Vasilyev, *J. Comput. Chem.* **2002**, *23*, 1254-1265.
[83] J.-L. Fattebert, F. Gygi, *J. Comput. Chem.* **2002**, *23*, 662-666.
[84] O. Andreussi, I. Dabo, N. Marzari, *J. Chem. Phys.* **2012**, *136*, 064102.
[85] A. J. Stone, M. Alderton, *Mol. Phys.* **1985**, *56*, 1047.
[86] S. L. Chan, C. Lim, *J. Phys. Chem.* **1994**, *98*, 692-695.
[87] S. Alvarez, *Dalton Trans.* **2013**, *42*, 8617.
[88] S. S. Batsanov, *Inorganic Materials* **2001**, *37*, 871-885.
[89] M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. A* **2009**, *113*, 5806-5812.
[90] J. Emsley, *The Elements*, Clarendon Press, Oxford, **1989**.
[91] M. Orozco, F. J. Luque, *J. Chem. Phys.* **1994**, *182*, 237-248.
[92] G. Hummer, L. R. Pratt, A. E. Garcia, *J. Phys. Chem.* **1996**, *100*, 1206-1215.